Visual Analytics Tool

for the Global Change Assessment Model

by

Zheng Chang

A Thesis Presented in Partial Fulfillment
of the Requirements for the Degree
Master of Science

Approved November 2015 by the
Graduate Supervisory Committee:

Ross Maciejewski, Chair
Hessam Sarjoughian
Dave White
Wei Luo

ARIZONA STATE UNIVERSITY

December 2015

ABSTRACT

The Global Change Assessment Model (GCAM) is an integrated assessment tool for exploring consequences and responses to global change. However, the current iteration of GCAM relies on NetCDF file outputs which need to be exported for visualization and analysis purposes. Such a requirement limits the uptake of this modeling platform for analysts that may wish to explore future scenarios. This work has focused on a web-based geovisual analytics interface for GCAM. Challenges of this work include enabling both domain expert and model experts to be able to functionally explore the model. Furthermore, scenario analysis has been widely applied in climate science to understand the impact of climate change on the future human environment. The inter-comparison of scenario analysis remains a big challenge in both the climate science and visualization communities. In a close collaboration with the Global Change Assessment Model team, I developed the first visual analytics interface for GCAM with a series of interactive functions to help users understand the simulated impact of climate change on sectors of the global economy, and at the same time allow them to explore inter comparison of scenario analysis with GCAM models. This tool implements a hierarchical clustering approach to allow inter-comparison and similarity analysis among multiple scenarios over space, time, and multiple attributes through a set of coordinated multiple views. After working with this tool, the scientists from the GCAM team agree that the geovisual analytics tool can facilitate scenario exploration and enable scientific insight gaining process into scenario comparison. To demonstrate my work, I present two case studies, one of them explores the potential impact that the China south-north water transportation project in the Yangtze River basin will have on projected water demands. The other case study using GCAM models demonstrates how the impact of spatial variations and scales on similarity analysis of climate scenarios varies at world, continental, and country scales.

To my loved ones for all the encouragement

ACKNOWLEDGMENTS

I would like to take this great opportunity to thank Professor Ross Maciejewski for his continued guidance and supervision in the past one and half year. I am thankful to Professor Hessam Sarjoughian, Professor Dave White and Dr. Wei Luo for being my master thesis committee members and for their support throughout the completion of my research.

TABLE OF CONTENTS

LIST OF FIGURES

Figure Page

Chapter 1

INTRODUCTION

This thesis presents the first ever web-based geovisual analytics tool for exploring the Global Change Assessment Model. The main focus is on enabling users to explore scenarios and perform inter-comparison and similarity analysis at different spatial scales and variations based on GCAM. The resulting patterns indicate that spatial scales and variations matter when analyzing similarity among different climate scenarios. This research sheds light on the inherently 'black box' structure in terms of modeling the complex relationships between climate change and human systems from a spatial scale perspective. For example, population growth and global climate change are two major stressors which have been identified as leading causes of concern for current fresh water supplies (Vörösmarty *et al.* (2010)). While future projections agree that there will be water scarcity, understanding the links between water scarcity and other sectors (such as energy, economy, etc.) can be difficult. These difficulties pose challenges for analysts and policy makers when exploring the impacts of potential future city growth and infrastructure projects. The major issue is that applying simulations to the scope of sustainability (including environment, economy, and society) requires the construction of large asynchronous simulation pipelines, where the output of simulation models becomes the input for one or more other simulations arranged in a sequence with feedback. Future sustainability issues need to be assessed in the context of climate change and climate mitigation policies (for example, estimating both water supply and water demand), and while models combining climate model outputs, water budgets, and socioeconomic information have been developed to understand future water scarcity (Vörösmarty *et al.* (2000)), it is difficult to understand their

strictly quantitative outputs. Integrating those models into a visual analytics interface will enable users to make sense of data, facilitate their hypothesis development, and support the decision-making process. As such, this thesis explores the design of the first ever web-based geovisual analytics approach for enabling the exploration of simulation runs from GCAM. This tool consists of geographical, temporal, and multidimensional spaces, which allows users to explore water scarcity in terms of geographical variations, temporal change, and scenario comparison with different future climate policies.

Understanding future climate conditions and their impacts is based, in part, on the ability to model the key drivers and underlying processes of complex interactions between climate change and human systems. Various climate scenarios can reveal different projections of future climate conditions and its cascading impacts. It is very important for analysts and policy makers to design adaptation and mitigation strategies for sustainability purposes based on different climate scenario results. Those climate scenarios can be modeled by the interaction between ensembles of climate models and human systems (e.g., water, energy use) with varied model parameters (Knutti (2008)). This assumes that those scenarios can approximately describe such interactions between climate models and human systems given limited observations, imperfect assumptions, and finite model choices (Knutti (2008), Beer *et al.* (2010), Yokohata *et al.* (2012)). Thus, it becomes necessary to investigate how inter-comparison and similarity analysis of climate scenarios influence variability in results. Yet, inter-comparison and similarity analysis of climate scenarios to understand the consensus among model results remains challenging because of the "black box" structure of modeling the complex interactions between climate change and human systems.

Research has begun exploring the reliability and quality of climate projections

through inter-comparison of ensemble climate model simulations with varied model parameters (Tebaldi *et al.* (2005), Furrer *et al.* (2007), Tebaldi and Knutti (2007)). The inter-comparison can help researchers gain insight into how model structures and varied parameters can impact the model outputs and how different models develop across institutes (Masson and Knutti (2011)) and evolve over time (Knutti *et al.* (2013)). Modeling climate and human systems is spatially dependent, but little research has explored the impact of spatial variations and scales on similarity analysis of climate scenarios. Thus, another goal of this thesis is to develop a novel geovisual analytics tool which allows users to explore the impact of spatial variations and scales on climate scenario comparison in terms of the interactions between ensembles of climate models and human systems with varied model parameters.

I have collaborated with climate scientists from the Joint Global Change Research Institute at Pacific Northwest National Laboratory, the developers of the Global Change Assessment Model (GCAM). GCAM is a global integrated assessment model combining representations of the global economic, energy, agricultural, land use, and climate systems (Clarke *et al.* (2007b), Kim *et al.* (2006)). It has made significant contributions to the Intergovernmental Panel on Climate Change (IPCC) climate change assessment reports (Clarke *et al.* (2007a)). Hejazi *et al.* (2014b) integrated water availability models into GCAM in order to quantify future water demand, supply, and scarcity in the context of the climate change. This thesis utilizes the new version of the GCAM model, integrating its input/output into a visual analytics tool to enable inter-comparison of scenario analysis on future water availability around the globe. The outputs of the GCAM model consist of water demand, supply, and scarcity at various spatial scales (e.g., global scale, regional scale), which allow us to explore the impact of spatial scales on scenario similarity. The development is motivated by the domain experts' need for understanding the impact of different scenarios.

Chapter 2

RELATED WORK

Given that an ensemble of climate models provides the basis for climate change projections, transparency in terms of model development and structure can contribute to model evaluation and its insight in terms of uncertainty and robust predictions (Pennell and Reichler (2011), Parker (2013)). Similarity and dissimilarity analysis of climate models/scenarios is an emerging research area which is helping analysts understand the impact of model structures on outputs. In this section I review what has been achieved in this area from both the climate modeling and visual analytics communities.

Previous work from visual analytics has focused primarily on the computational steering side of simulations. Computations steering is a commonly used concept that allow users to change parameters of their simulation during its execution and get the feedback immediately. This concept is popular since simulations become more complex, have many different input parameters and large amounts of heterogeneous data results. For example, Waser *et al.* (2010) proposed World Lines and Visdom (Ribicic *et al.* (2013)) to help users manage ensemble simulations of complex flooding scenarios. They also extended their work to support interactive comparison between the original simulation output and its alternative after steering process. For my research, I argue that there needs to be more focus on the policy side of the visualization component. Essentially, there is a large potential for applying visual analytics to complex systems that involves both the efficient presentation of simulation results to analysts and to policy planners. In this section, I will also discuss previous work in visualizing the modeling and simulation domain and climate visualization.

2.1 Clustering and Climate Models

Similarity analysis of ensembles of climate models can help users gain insight into the reliability of model outputs. Researchers applied clustering approaches to quantify the differences between climate models in order to understand the impact of the model structures on the output. For example, Pennell and Reichler (2011) applied hierarchical clustering to 24 climate models that simulate the observed present-day mean climate and found that similarities in model implementation determine ensemble estimation. Masson and Knutti (2011) performed a similar analysis using a hierarchical clustering approach to the entire spatial field of monthly temperature and precipitation values from multiple climate models across different ensembles and found that models from the same institution tend to exhibit strong similarities. Huntzinger *et al.* (2013) also applied a similar approach to terrestrial biosphere models (TBMs) to highlight the impact of model similarities and dissimilarities on carbon cycle, vegetation, energy, and nitrogen cycle dynamics. Knutti *et al.* (2013) further argued that the similarities are caused by the interdependence of most models on their predecessors, which makes the interpretation of multimodel ensembles more complicated.

The previous research demonstrates the impact of climate model structures on outputs with varied parameters without considering the impact of spatial scales and variations. Modeling interactions between climate models and human systems is spatially dependent, as sum, this thesis investigates the impact of spatial scales and variations on different climate scenario outputs in this paper.

2.2 Modeling and Simulation Visualization

To address sustainability issues, we must handle systems of systems where data in each simulation module consists of different spatial and temporal resolutions, the output from the models need to be explored, compared, and cross-correlated, and, most importantly, the interactions between models must also be understood. Traditionally, visualization of simulation runs focuses on the domain expert analyst or model builder. For example, Matković *et al.* (2010), their work implemented interactive visualization functions allow users to focus on interesting cases, which can reduce overall simulation runs. Maciejewski *et al.* (2011) applied visual analytics techniques to support the exploration of parameter changes during disease simulation spreads, and Andrienko and Andrienko (2013) proposed a comprehensive visual analytics environment for spatial modeling libraries. The IBM Supply Chain Simulator (SCS) (Buckley and An (2005)) which visualizes relationships in supply chain. Hubmann-Haidvogel *et al.* (2012) developed a visual analytics tool to deal with massive social media data from different resources. They extended their original work (Hubmann-Haidvogel *et al.* (2009)) with three new visualization techniques: dynamic topography information landscapes, news flow diagrams and longitudinal cross-media analysis to support dynamic character exploration. To overcome the data accessing problem in urban research. Tomko *et al.* (2012) developed a web-based visual analytics tool allows researchers to access federated urban data which can be further visualized and analyzed online. They implemented RESTFul services and multi-resolution map-based analysis features to support exploratory data analysis. Kohlhammer *et al.* (2012) defined three major tasks for policy making: information foraging, policy design, and impact analysis. They also categorized a series of visualization disciplines including information design, information visualization, semantics visualization, vi-

sual analytics (VA), and knowledge discovery and data mining (KDD), for policy making. Malik *et al.* (2014) presented a predictive visual analytics environment in supporting decision-making in terms of effective resource allocation and deployment through allowing users to focus on appropriate spatiotemporal resolution levels. Luo (2014) applied a visual analytics approach to support the design of disease control strategies at appropriate spatial resolutions in the epidemic domain. Kothur *et al.* (2014) presented a visual analytics tool to compare ocean model output with reference data in order to detect and analyze geophysical processes in terms of their differences and similarities. Houghton *et al.* (2012) proposed a web-based visualization tool, GEMSS (Geospatial Emergency Management Support System), for public officials in Austin, Texas to explore heat-health related data. Poco *et al.* (2014b) proposed a visual analytics tool that allows users to design and explore grouping data, to determine potential relationships between criteria sets and spatiotemporal behaviors. Their work allows users to design data groups for weights optimization. Different visualization techniques such as timeline and matrix view are commonly used as linked views to present different steps of a workflow. However, their work has limited geospatial analysis ability which is considered an important approach for climate data. Sanyal et al. introduced Noodles to explore weather prediction related uncertainty. They introduced glyphs, ribbons, and spaghetti plots with a grid view for numerical weather models and uncertainty visualization in Sanyal *et al.* (2010), concentric circular glyphs provide information about the deviation of ensemble members at a given grid location. Overlapped ribbons along a given contour represents different ensemble member's deviation since the ribbon segments' width is determined by uncertainty at that point. Finally, the spaghetti plot is implemented to interactively show local ensemble mean or bootstrap mean. Steed *et al.* (2013) implemented a visual analytics system called the Exploratory Data analysis Environment (EDEN)

which can be used to explore complex Earth system data sets. In their work, filtered model data will be visualized using a parallel coordinate plot(PCP) based technique to provide exploratory data analysis. Buja *et al.* (1996) proposed a taxonomy for interactive visual analytics tool based on different purposes. They also introduced XGobi for multivariate data exploration consists of various techniques such as focusing, linking, and arranging views.

2.3 Climate Visualization

Visualization is considered to be a key technology for representing and analyzing climate models and observations as well as relevant social and environmental data (Nocke *et al.* (2008)). Designing intuitive visual representations in the climate context to facilitate hypothesis development by identifying complex patterns is an active research area. For example, Kehrer *et al.* (2008) demonstrated how interactive visual exploration can facilitate hypothesis generation with climate data. Ladstädter *et al.* (2010) demonstrated that the integration of classical statistics and interactive visual exploration can enhance the workflow of climate analytics. Jin and Guo (2009) presented a multivariate geovisualization tool to support discovery and understanding of unknown complex patterns from the heterogeneity of climate related data, including space, time, and multiple variables. Williams *et al.* (2014) developed UV-CDAT to interactively analyze and visualize large-scale climate data. Wang *et al.* (2008) focused on volumetric time-varying visualization which is used to select time steps that contain major features. Sun *et al.* (2012) proposed a web-based visual analysis platform to manage and explore distributed 3D climate data. Nocke *et al.* (2007) proposed a visualization library with SimEnvVis framework to explore and evaluate climate simulation outputs. Burch *et al.* (2010) proposed a new workflow for visualizing local climate change data using iteratively generated 3D images. Dockerty *et al.*

(2005) also developed a GIS-based visualization tool for presenting climate change impacts at the local or landscape scale. As an extension, they use rendered 3D images to visualize predicted land use changes. Helbig *et al.* (2014) developed a workflow for raw climate simulation data visualization and the result can be explored both on the desktop and in a virtual reality environment. For different types of variables, Helbig et at. applied a series of visualization techniques such as texture, iso lines, iso surfaces, and streamlines. Interactive functions are also implemented that allow users to reduce number of variables and size of sample data. To explore potential relationships between climate change and energy use, Ismail *et al.* (2014) integrated choropleth maps and bivariate maps in a geographic information system (GIS).

In addition to supporting pattern identification and hypothesis generation, the evaluation of climate models through visual analytics is another major research direction. For example, Steed *et al.* (2012) evaluated climate models by identifying relationships between climate variables and their geographic associations. Their focus is facilitating the assessment in a single climate model and analysis of the interdependence among variables with respect to the amount of variability which can be attributed to specific other variables. To address this challenge, they introduced a parallel coordinate plot (PCP) based visual analytics tool named EDEN and applied it to evaluate the Community Land Model Version 4 (CLM4). However, their tool has limited abilities for simulation (model) comparison, which I argue that there needs to be more focus on. Potter *et al.* (2009a) integrated ensemble datasets that combine multiple numeric climate models into the ViSUS/Climate Data and Analysis Tools (CDAT), in order to support uncertainty exploration. In their work, they computed the mean and standard deviation for every 3D spatial location with the value retrieved from the ensemble members. After computation, they plot the result using a series of visualization techniques such as height fields and iso-contours. Potter *et al.*

(2009b) also developed Ensemble-Vis, a framework that can be used to explore and analyze ensemble data by providing users statistical visualizations. They argued that linked views allow users to gain insight into the data by exploring simulation result distribution and model uncertainty. Dasgupta *et al.* (2015) evaluated the use of different climate data visualizations (e.g., scatterplot, map) for climate model comparison. Brus *et al.* (2013) introduced several uncertainty visualization methods combined with interpolation functions for climatological data analysis. Their work compared interpolation approaches based on commonly used uncertainty visualization techniques in GIS analysis. By comparing existing intrinsic and extrinsic approaches, Kaye *et al.* (2012) suggested 6 guidelines for developing appropriate methods for mapping climate variables, including uncertainty. Poco *et al.* (2014a) developed SimilarityExplorer, an exploratory visualization tool to support fast similarity analysis. Researchers are able to gain an overview of model similarity and uncertainty in terms of geographical variations and temporal change. However, their tool lacks abilities that allow users to focus on specific geographic areas for further model similarity evaluation.

Chapter 3

SYSTEM

My work aims to develop visual analytics to support both model exploration and comparison in the context of climate change in order to address the analysis of complex modeling tools with a focus on how to enable simulation comparison for advanced analysis. I propose a novel web-based tool which allows users to explore GCAM water data including demand, supply, and scarcity in terms of geographical variations and temporal change. The tool also enables users to formulate "what-if" questions by building different scenarios with different future climate policies. Besides, I also would like to apply clustering approach to explore inter-comparison between climate scenarios with the consideration of the impact of spatial variations and scales. Therefore, I applied visual analytics approaches to address the following high-level analysis questions: 1) how does modeling interactions between climate change and human systems with a varied combination of parameters cause similarity or differences in model outputs; 2) how do spatial variations and scales impact scenario analysis in terms of similarities or differences in model outputs.

3.1 Data Structure

In this study, I applied the new version of the GCAM model which has explicitly incorporated water demand, water supply, and water scarcity (Hejazi *et al.* (2013a), Hejazi *et al.* (2013b), Hejazi *et al.* (2014a)). I consider sixty different scenarios which include three future precipitations, two future emissions, five global populations, and two China South-North Water Transportation Projects. The future precipitation levels are based on three Global Climate Models (GCMs) which correspond to water

availability at dry, mild, and wet futures, respectively. The two future emission levels include two Greenhouse Gas (GHG) control policies indicating moderate climate change and low climate change. The five global populations range from 8 to 10 billion in steps of 0.5 billion. The China South-North Water Transportation Project from the Yangtze River Basin to the Ziya He Basin allows users to understand its impact on both basins. GCAM estimates water demand, water supply, and water scarcity at two different spatial scales: the water region scale with 62 regions in total (including 31 divisions from China) and the water basin scale with 235 basins in total at five year intervals (2010 to 2095). Water demand includes demand from irrigation, demand from demand from livestock, demand from electricity, demand from manufacturing, demand from domestic, and primary energy production. All sixty scenarios include 7 variables (e.g., water demand, supply, and scarcity), 235 water basins/62 water regions, and 18 single years at five year intervals (2010 to 2095). Within each scenario, the basin similarity includes 7 variables and 18 single years at five year intervals (2010 to 2095).

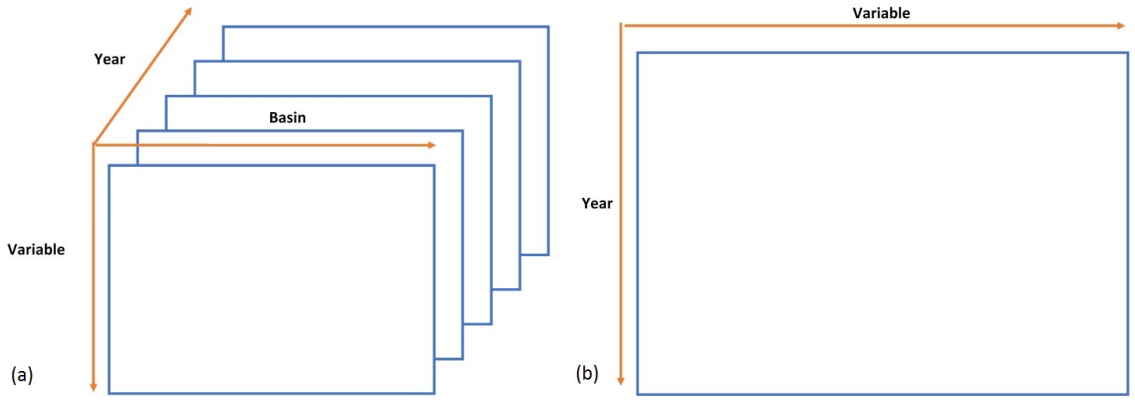

Figure 3.1: The data structure on which the calculation of scenario and basin similarity is based. (a) The data structure for the scenario similarity. (b) The data structure for the basin similarity.

All sixty scenarios can be conceptually constructed as a three dimensional matrix (Figure 3.1 (a)), including 7 variables (water demand from irrigation, water demand from water demand from livestock, water demand from electricity, water demand from manufacturing, water demand from domestic, water supply, and water scarcity), 235 water basins or 62 regions, and 18 single years. Within each scenario, I also would like to explore the basin similarity according to 7 variables and 18 single years. Then the data structure is represented as the two dimensional matrix (Figure 3.1 (b)).

3.2 Shapefile

The shapefile is developed and regualted by ESRI as an open standard format for storing geometric location and correlated attribute information (ESRI (1998)). Although this format has limited capacity to store topological information, it is widely used because it is simple and can store the primitive geometric data types such as points, lines, and polygons together with attribute data. Various open source APIs support shapefiles as interactive layers for map views, such as Leaflet. I used shapefiles to store boundaries, names, and selected values for basins and regions (Figure 3.2).

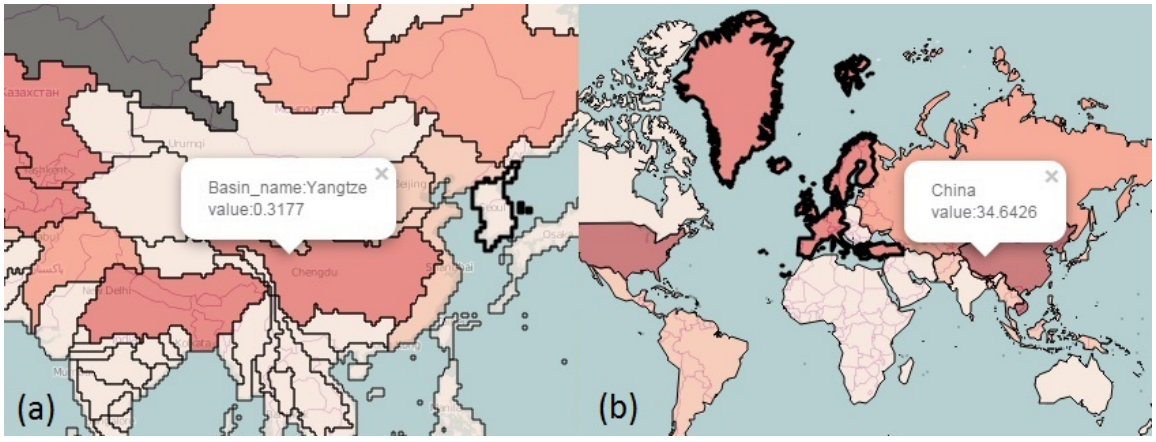

Figure 3.2: In this system, I store the basin/region names ((a) and (b)) and attribute values. For example, after users select one of the six water demand attributes (manufacturer, irrigation, livestock, electricity, domestic, and total), the shapefile will be updated and then be plotted on the map (a).

3.3 Hierarchical Clustering Approach

I applied a hierarchical clustering algorithm (Pang-Ning *et al.* (2006)) to analyze similarity between scenarios/basins. The algorithm is the most popular approach to understand the impact of model structures on the output. I use an agglomerative strategy to merge clusters, Euclidean distance (Deza and Deza (2009)) to represent model similarity, and average linkage clustering (Szekely and Rizzo (2005)) as linkage criteria. The workflow shows as below:

- Calculate Euclidean distance matrix between all scenarios/basins.
- The distance between two scenarios (MT, MS) is computed by the equation:

$$Dist(MT, MS) = \sum_{i \in Y} \sum_{j \in V} \sqrt{\sum_{k \in B} (MT_{i,j,k} - MS_{i,j,k})^2} \tag{3.1}$$

  Where $MT_{i,j,k}$ represents a value retrieved from scenario MT in the i^{th} year in year set Y at five year intervals (2010 to 2095), the j^{th} variable in variable set V of 7 water variables, and the k^{th} basin in the selected basin set B.

- Distance between two basins (BT, BS) within one scenario is computed as:

$$Dist(BT, BS) = \sum_{i \in Y} \sqrt{\sum_{j \in V} (BT_{i,j} - BS_{i,j})^2} \tag{3.2}$$

  Where $BT_{i,j}$ represents a value retrieved from the basin BT in the i^{th} year from the year set Y at five year intervals (2010 to 2095), and the j^{th} variable from the set V of 7 water variables.

- Merging the two closest clusters into one new cluster based on the average linkage equation:

$$Dist(A, B) = \frac{1}{|A|\,|B|} \sum_{a \in A} \sum_{b \in B} dist(a, b) \tag{3.3}$$

  Where A and B are clusters and dist is the distance matrix.

- Continue step 2 and step 3 until there is only one cluster.

3.4 System Design

This system provides a web-based visualization tool for GCAM which explores water-related climate change scenarios. GCAM is a dynamic-recursive model including representations of the global economy, the energy system, agriculture and land use, and climate (Kim *et al.* (2006), Clarke *et al.* (2007a)). Hejazi *et al.* (2014b) explicitly incorporated water demand, water supply, and water scarcity in GCAM. Water demand includes the following categories: irrigation, livestock, electricity, manufacturing, domestic, and primary energy production. GCAM tracks water demand, water supply, and water scarcity at different spatial scales, including 14 geopolitical regions, 235 water basins, and a grid scale of ($0.5°\times0.5°$). GCAM estimates annual water demand, supply, and scarcity every five years from 2010 to 2095.

The GCAM geovisual analytics tool allows analysts to explore water supply, water demand, and water scarcity in the context of climate change and climate policies. I also make use of choropleth maps to support geographical analysis as well the component coordination methods to enable dynamic linking and brushing across views. I add a dendrogram component to support the exploration of the impact of spatial scales and variations on scenario similarity. I also add a timeline view to support the exploration of the average scarcity level of different scenarios over space and time. And a parallel coordinate plot to support the exploration of basin similarity within one scenario over space.

The system is developed as a Modelviewcontroller (MVC) architecture. On the model side, I implemented several RESTful services with JAVA for retrieving data from database and responding requests from clients. On the controller side, I implemented functions using JavaScript to call backend RESTful services through AJAX

library to retrieve data. For the view side development, I used the D3.js library, OpenStreetMap, and other open source plugins to visualize water data.

3.4.1 MVC Framework

Model-View-Controller is a software architectural pattern which is widely used in UI development. It separates the different aspects of applications while providing a loose coupling between them (Reenskaug and Coplien (2009), Burbeck (1992)). In this system, I implemented RESTful services using JAVA for data retrieval and pre-processing as model components. For visualization, I developed visual components using JavaScript and open-source libraries such as D3.js (Bostock (2012)) and Leaflet. For the controller, I designed combo-boxes, sliders, and interactive functions which allow users to update the model's state or change the presentation of the model (Figure 3.3).

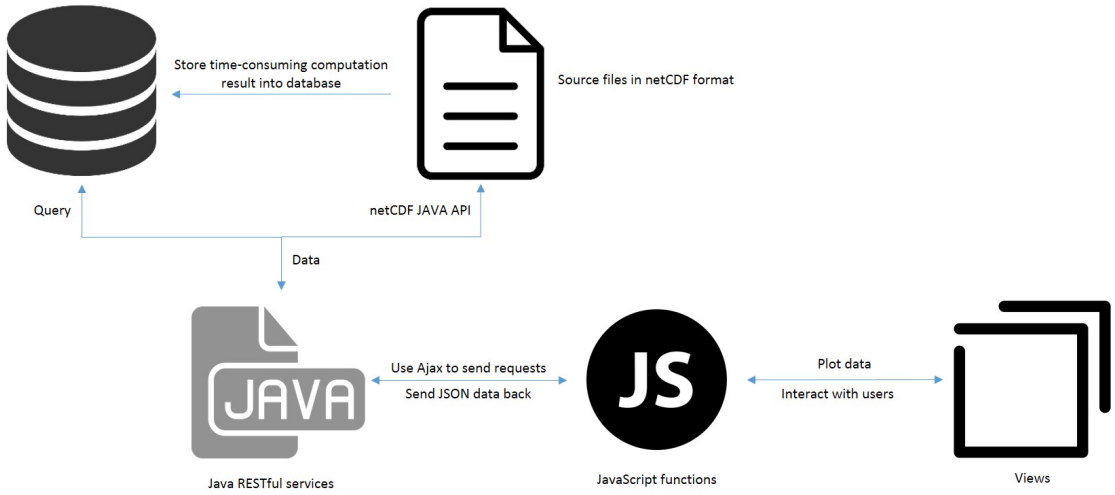

Figure 3.3: The system is designed under the Model–View–Controller (MVC) framework. For view, maps, charts, timelines, dendrograms, and PCP (Parallel Coordinate Plot) are implemented as visual components. On the model side, RESTful services are used to retrieve and pre–process data from the database and source file. Combo–boxes, sliders, and interactive functions are major controller components. The data is passed between controller and model, model and view using JSON objects through AJAX requests.

3.4.2 Visual Components

The visualization design follows the Visual lnformation–Seeking Mantra: overview first, zoom and filter, then details–on–demand (Shneiderman (1996)). The system includes two sets of linked views. The first view contains a line chart and three synchronized maps. This view provides the general view of the scenarios. The second view includes a map, dendrogram, parallel coordinate plot, and labeled time line, which compares the similarities between scenarios. Various operations are provided to perform the tasks such as synchronized focusing, area selecting, filtering and brushing.

Line Chart

There are two types of line charts that have been applied to out tool: single line chart and multi–line chart. The single line chart is used in the GCAM visual analytics view to present a summary value of water demand attributes of selected scenarios by year. The x–axis is labeled with the year tag from 2010 to 2095 in steps of 5 years and the y–axis is labeled with value range of selected attribute. Users can switch between different water demand attributes by clicking on buttons above the single line chart (3.4). Water demand map in GCAM visual analytics view will also be updated after user changes the water demand attribute (Figure 3.4).

The multi–line chart is designed to explore how the China South–North Water Transportation Project will affect the water supply and water scarcity of Yangtze River Basin and the Ziya He Basin under different climate scenarios. The design of the axis is the same as the single line chart. There are three buttons stand for future precipitation, emission, and global population respectively, and each of them has a combobox next to it. By adjusting values in the three comboboxes, users can select models with the same precipitation, emission and global population but

different water project implementation (on and off). By clicking on a button, users can select models with all possible values of the selected parameter, different water project implementation, and same values for others. By adjusting the three model features using buttons and combo-boxes together, users can form scenario sub-groups for detailed exploration (Figure 3.5).

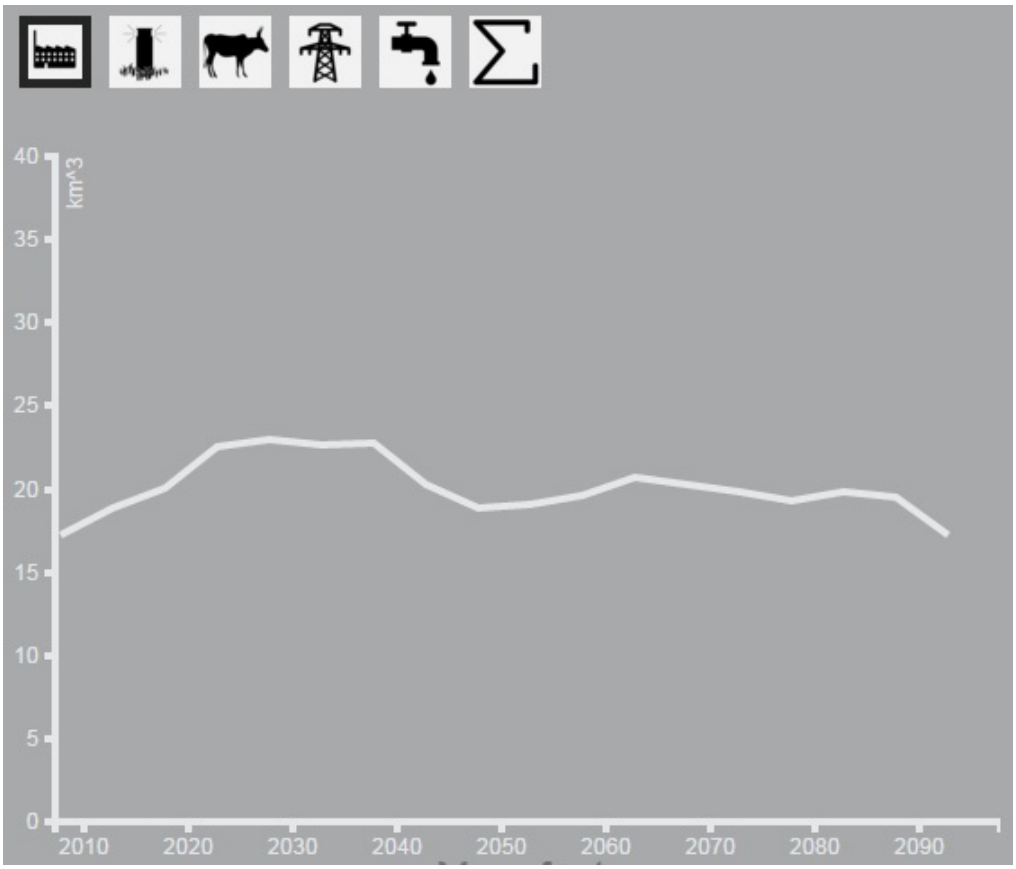

Figure 3.4: A single line chart for the yearly summary of water demand. Highlighting the boundary for selected button and mouseover tooltips helps users gain a better understanding.

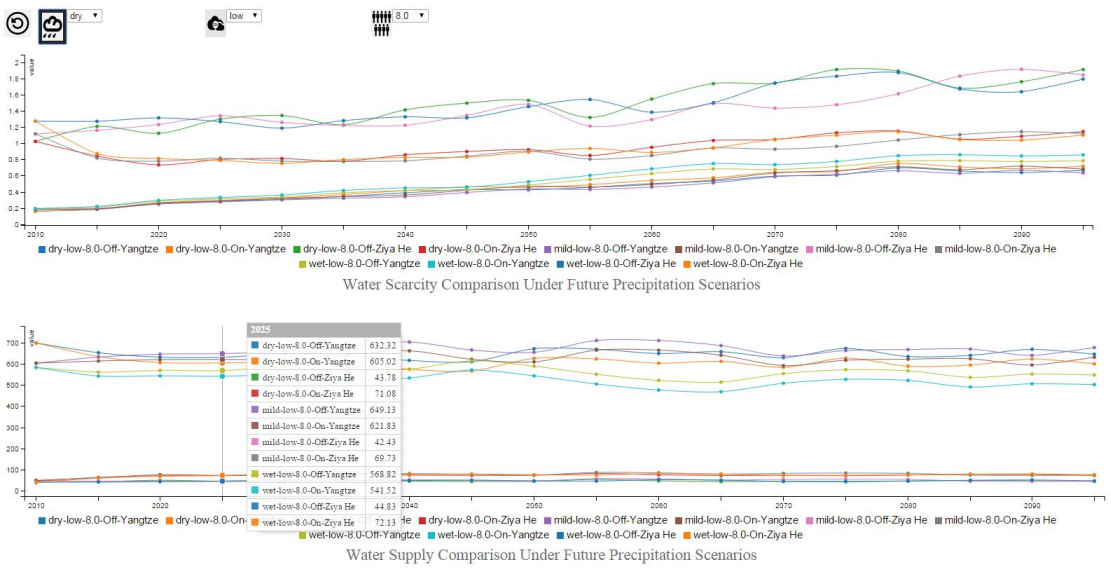

Figure 3.5: Popup tables present the precise values of all plotted scenarios in that year, and legends are provided under the charts to help users match the plotted lines to scenario names.

Map Plot

Champlain maps are one of the most commonly used techniques for geographical data visualization. This work implements Champlain map as the major visual component for scenario exploration and similarity comparison. To achieve these goals, interactive functions such as drawing and focus-plus-context are developed. Color schemes and legends are also added to the map plots in order to link all the other visual components together to assist users in scenario exploration and similarity comparison.

In the GCAM visual analytics view, three synchronized maps show the yearly water data of a selected scenario from three different perspectives: water demand, water supply, and water scarcity. To assist users in gaining insight into some specific areas, I implemented synchronized maps so that when users change the focus of any one of them by dragging or zooming, the others will update their focus at the same center location and zoom level. This feature makes sure that users will always focus on their area of interest and facilitate their exploration. At basin and region level,

users can also explore names and precise values by clicking on basins or regions of interest(Figure 3.6).

In the similarity analysis view, maps are integrated with more interactive features and are linked to dendrograms, parallel coordinate plots, and labeled time lines. I implemented rectangle and polygon selection functions, which allow users to form basin sub-groups and apply them to scenario similarity analysis and term exploration. Users can draw either polygons or rectangles on the map to select an area of interest, after that, basin sub-groups will be created and passed to the clustering process and time line filtering process. The editing function for polygon and rectangle selection allows users to adjust their area of interest by dragging or adding vertices, and moving the entire object (Figure 3.7).

Dendrogram

The dendrogram is one of the most frequent visualization approaches used to illustrate the arrangement of the clusters in hierarchical structure (Everitt and Skrondal (2002)). My tool implements a dendrogram view to illustrate the hierarchical clustering results for scenario similarity. Leaf nodes represent the scenarios and parent nodes represent the grouping procedure of the hierarchical clustering process. To help users gain a better understanding of the clustering results, I added legends to match the parent nodes with correlated model variables (Figure 3.8). Furthermore, different clicking functions such as filtering and basin detail exploration on parent nodes and leaf nodes are implemented for linking the dendrogram with a parallel coordinate plot and a labeled time line.

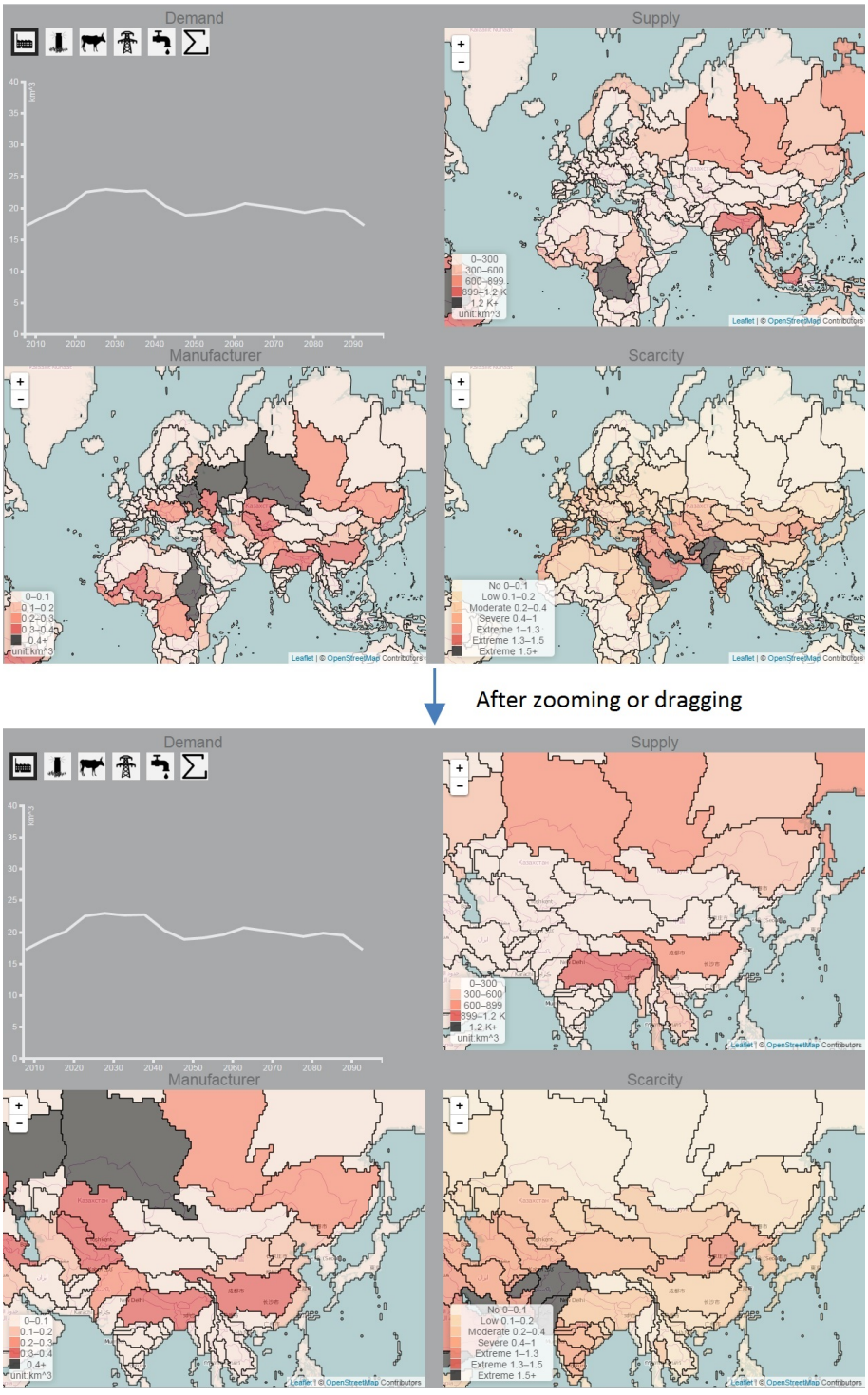

Figure 3.6: Synchronized maps. After zooming and dragging on the bottom left map, the other two maps update their focuses automatically. The legends clearly present the overall data distribution and basin/region details.

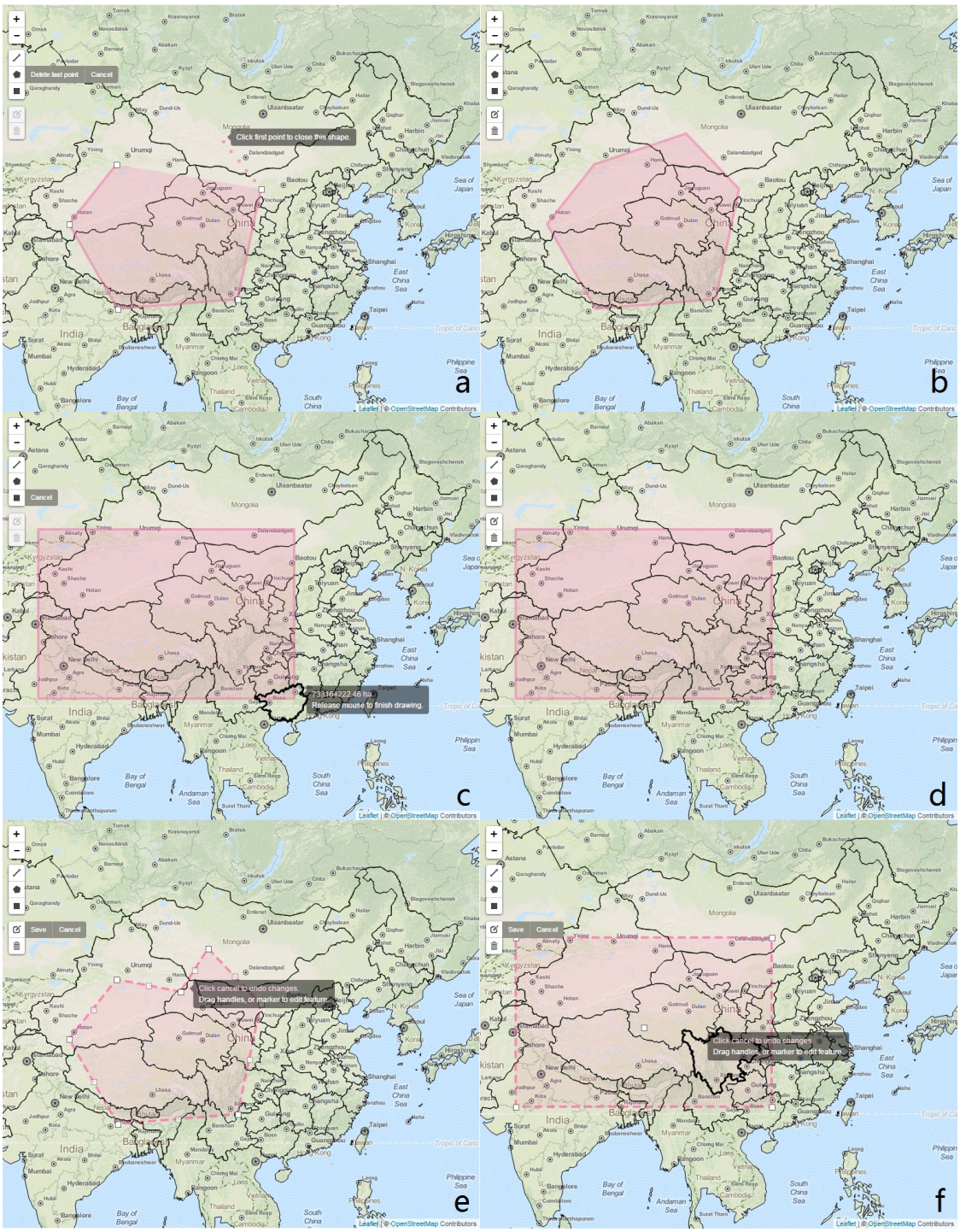

Figure 3.7: Click on "draw polygon" or "draw rectangle" to activate the area selection tool. For polygon drawing, click to add vertices and close the shape by clicking the first vertex (a)(b). For rectangle drawing, click and drag (c)(d). To edit a polygon, click on the edges to add new vertices and drag existing vertices to new position (e). For editing rectangle, drag existing vertices to resize, click on center point and drag to move (f).

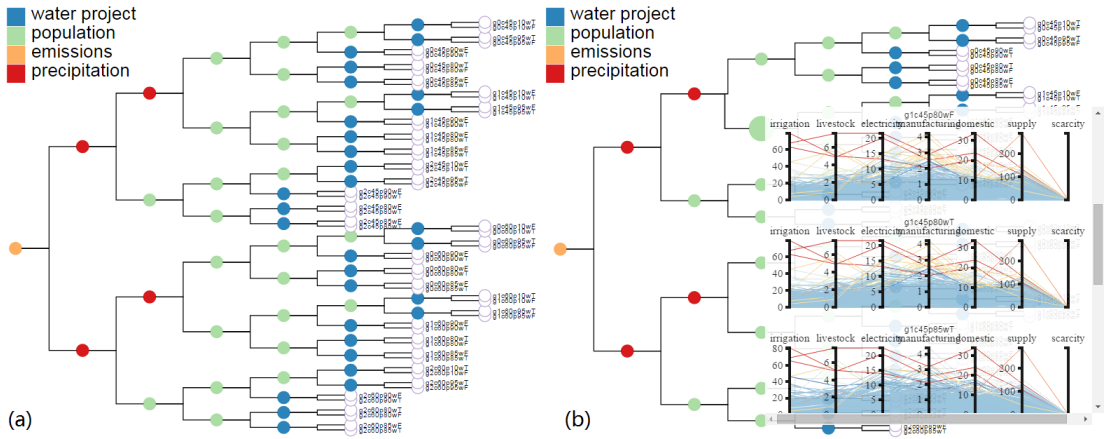

Figure 3.8: Dendrogram in similarity analysis view (a). Users can filter the models shown on labeled time line by clicking on parent nodes. After filtering process, only models belong to the selected sub-tree will be displayed. The root of the selected sub-tree will be enlarged to indicate users' selections (b). An overview of basin level similarity for all sub-tree models is then displayed. Clicking on leaf nodes will enable similarity comparison on the basin level within the selected scenario.

Parallel Coordinate Plot

Along with model level similarity analysis, my tool also provides within model basin level similarity comparison. In this work, I focus on the selected sub-group formed by the selection function on the map view, and I visualize the clustering results using a PCP. After area selection and model selection, my tool calculates the similarity between basins based on their water supply, water scarcity, and five water demands (irrigation, livestock, electricity, manufacturing, and domestic usage). I synchronized the color scheme on the parallel coordinate plot and the map view to help users gain a better understanding of the clustering results from both the geospatial perspective and the data distribution perspective (Figure 3.9). Basins belonging to the same cluster will have the same color, and for each basin, its color is synchronized on both the PCP and Champlain map view. In this way users can explore the clustering result from the geospatial perspective by looking at the map while having an intuitive sense of the data distribution from the parallel coordinate plot (Figure 3.9).

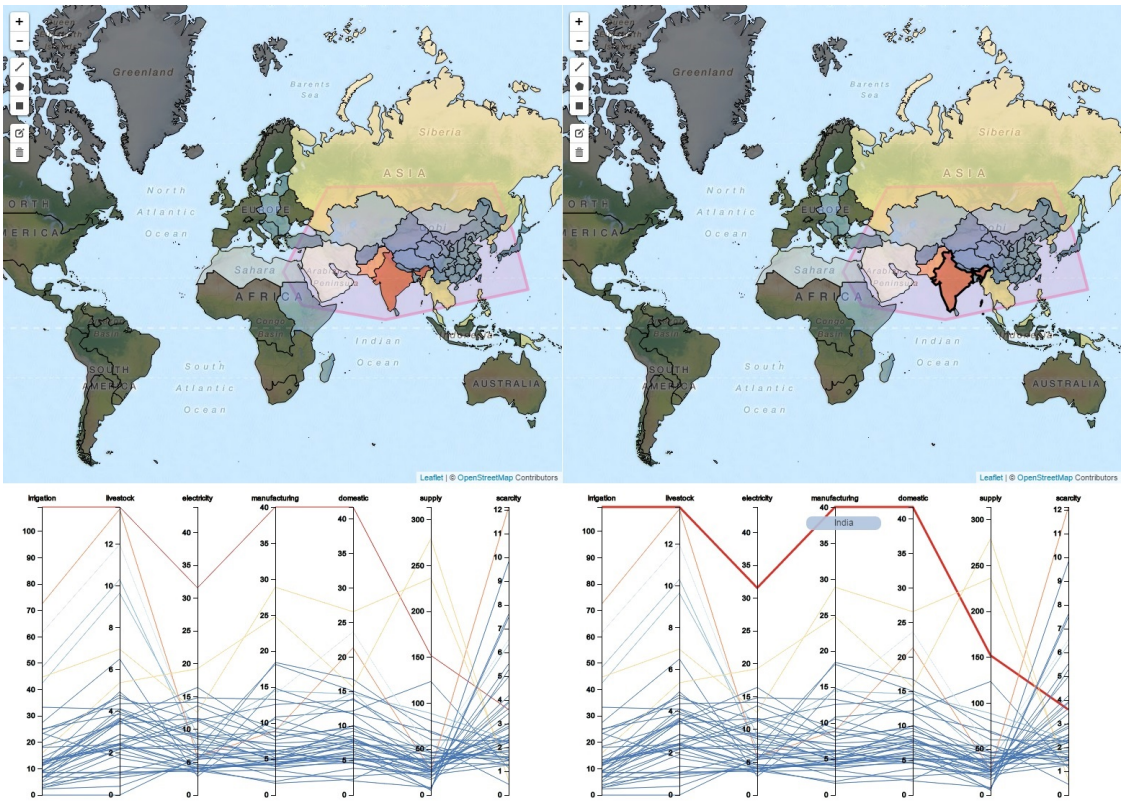

Figure 3.9: Parallel coordinate plot with a synchronized color scheme. Mouseover tooltips and highlighting on the PCP help users locate the basins even faster.

For further analysis on basin similarity, my tool implements a brushing function n the PCP's axes. By brushing an axis, users can select a basin sub-group of the previous selected area and rerun the clustering algorithm. The results are visualized in the same way except those basins have not been brushed and their color will be restored (Figure 3.10).

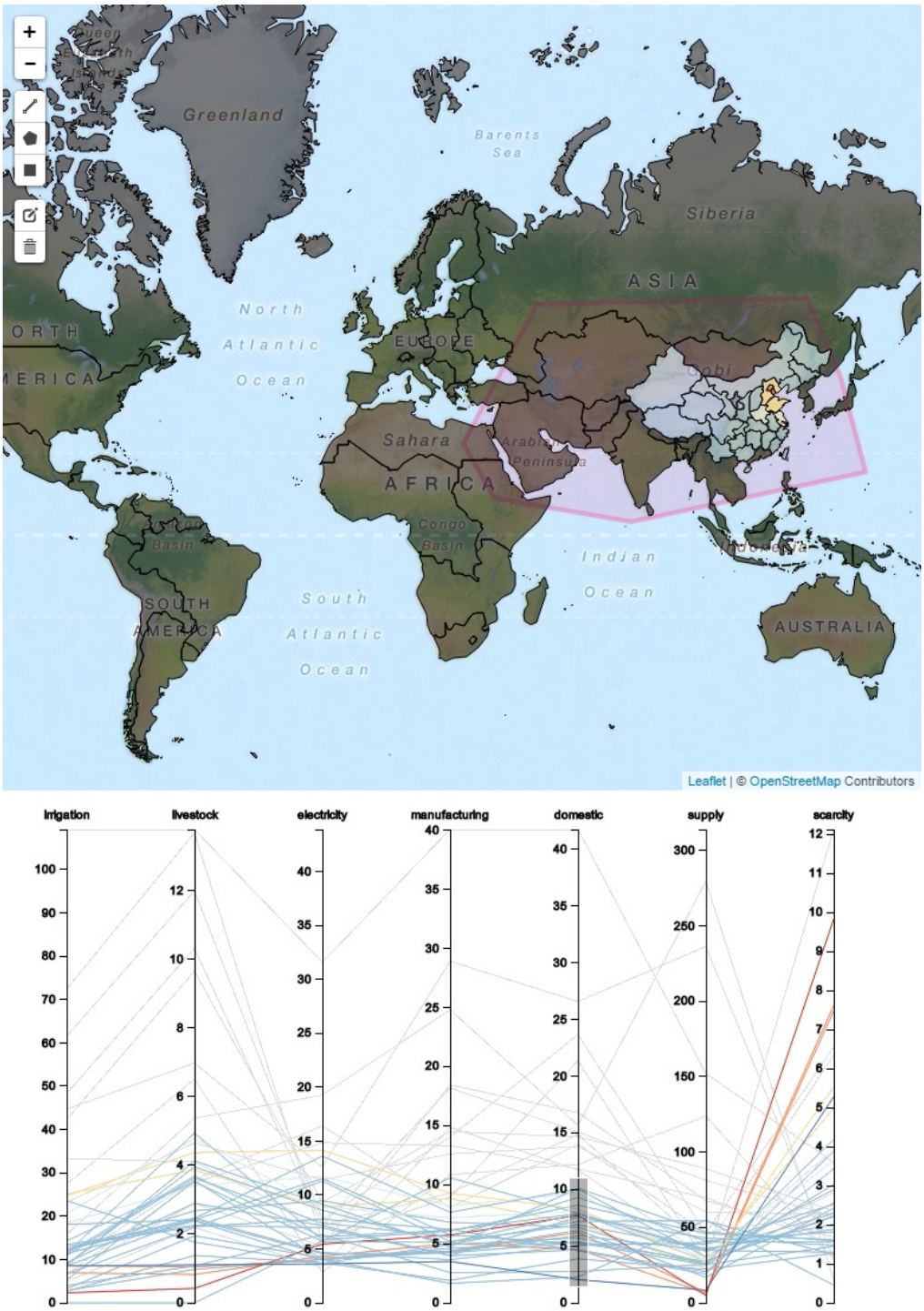

Figure 3.10: After brushing on an axis, the clustering process will rerun on the selected basin sub-group. For example, after brushing on "domestic" axis on PCP, the unselected basins will fade on both PCP and map view. The updated result will be visualized using the same strategy.

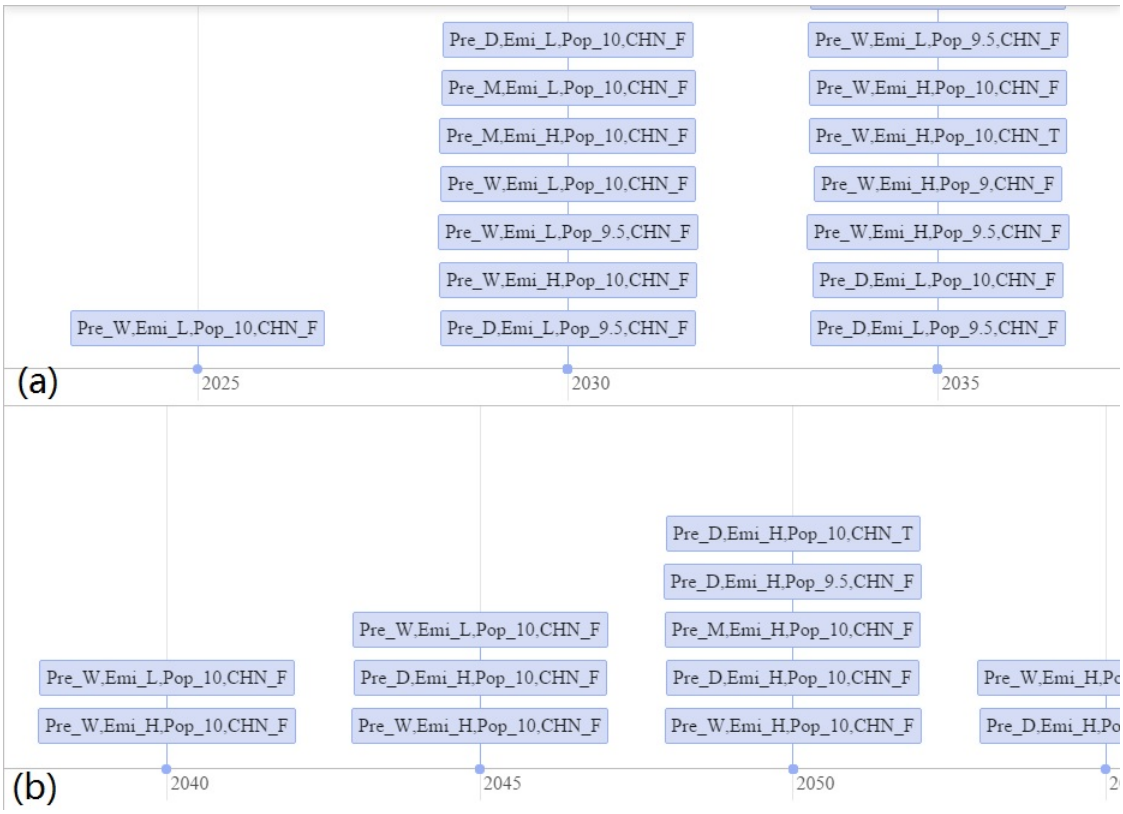

Figure 3.11: The original result based on the selected area with a threshold of 0.5 (a). The filtered result with a threshold of 0.54 (b).

Labeled Time Line

One of my novel approaches is exploring water scarcity in terms of temporal change and scenario comparison. To achieve this goal, I introduced a labeled time line which is used to show how the climate changes and water policies will affect water scarcity by year. In the dendrogram, I implemented a filtering function to help users reduce the total number of scenarios displayed on the labeled time line. By clicking on parent nodes in the dendrogram, users can choose a sub-tree, then only scenarios belonging to it will be passed to the labeled time line for the water scarcity evaluation process. After the filtering process, my tool will compute the yearly average of water scarcity for each selected scenario. For basin level similarity comparison, I only use basins

belonging to the selected area in the map view as the representatives of the scenarios. In the country level similarity comparison, my tool allows users to choose only one basin on the map as the representative of the scenarios. For each year, if a scenario's water scarcity average at this year is higher than the threshold set by the users, a label with this scenario's name will be added to the time line. (Figure 3.11) With the help of the labeled time line, users can easily explore the scenario's water scarcity by time and gain insight into the impacts of climate change and water policies on specific areas.

3.5 Views

To address the two purposes (scenario exploration and similarity comparison), I designed and developed two views. I implemented GCAM visual analytics view for scenario exploration which helps users gain an overview of all the 60 different models and analyze the impact of the China South-North Water Transportation Project on the future water scarcity of its donor basin (Yangtze River) and recipient basin (Ziya He). After users get the "big picture" of all the 60 different models, the similarity analysis view provides users a hierarchical clustering based comparison approach, which allow users to analyze similarity on both model and basin levels.

3.5.1 GCAM Visual Analytics View

The GCAM visual analytics view consists of three sub-views: the control scenario design view (Figure 3.12), the spatiotemporal view (Figure 3.12), and the scenario comparison view (Figure 3.5). The control scenario design view allows users to design different scenarios by combining four features: future precipitation, emission, global population, and the China South-North Water Project. The future precipitation features allow users to pick among three global climate models to calculate

water availability at dry, mild, and wet futures. The emission feature allows users to design two mitigation strategies in terms of greenhouse gas emissions at low and high levels. The global population feature allows users to estimate the global population in 2050 from 8 to 10 billion in steps of 0.5 billion. The China South-North Water Transportation Project from the Yangtze River Basin to the Ziya He Basin allows users to turn on or off the project to understand its impact on both basins. In total, for the three future precipitations, two future emissions, five global populations, and two China South-North Water Projects, 60 (3×2×5×2) different combinations are simulated for user exploration.

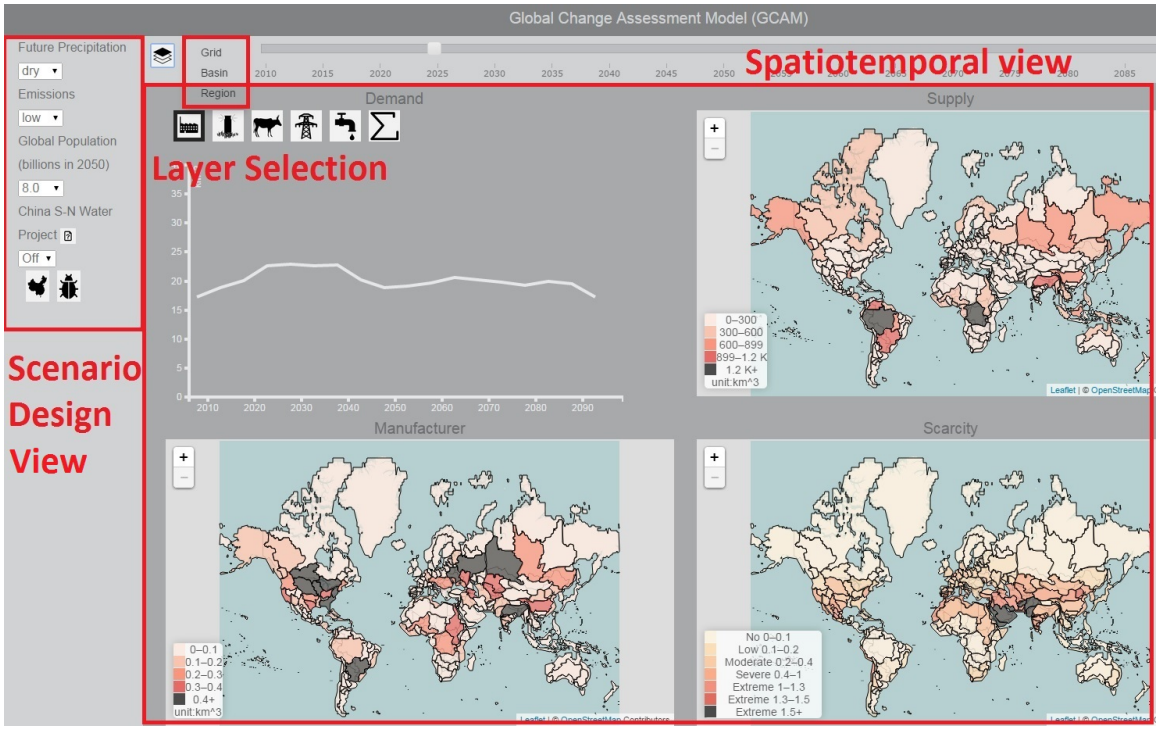

Figure 3.12: The GCAM visual analytics tool at the basin scale. The legend design for water scarcity index (WSI) is adopted from Raskins definition Foley *et al.* (1994) : no scarcity ($0 \sim 0.1$), low scarcity ($0.1 \sim 0.2$), moderate scarcity ($0.2 \sim 0.4$),severe scarcity ($0.4 \sim 1$), and extreme scarcity (> 1). The scenario design panel is on the left, the layer selection allows the map to switch between grid scale, basin scale, and region scale, and the time slider on the top allows users to explore different scenarios from the temporal perspective.

The spatiotemporal view allows users to explore water supply, water demand, and water scarcity at three geographical scales: the grid scale at ($0.5°\times0.5°$) resolution, the water basin scale with 235 different areas, and the geopolitical region scale with 14 regions every five years from 2010 to 2095. Three geographical maps are used to explore water supply, water demand, and water scarcity respectively. Water demand includes six different demand types: manufacture, irrigation, livestock, electricity, domestic and total. The three maps support coordination, so users can explore the same geographical locations from an integrated perspective of water supply, demand, and scarcity. The scenario comparison view allows users to compare the Yangtze River Basin and the Ziya He Basin in terms of their water scarcity over time. Users can compare the China South-North Water Transportation Project under different scenarios in terms of three future precipitations, two greenhouse gas emissions, and five global populations.

After scenario selection has been done using the combo-boxes in the control scenario design view, the overview of the simulation result will be presented by the line chart and three Champlain maps in spatiotemporal view. The line chart on the top-left shows the world summary of water demand in six different measures (manufacturer, irrigation, livestock, electricity, domestic, and total) by year. The map on the bottom-left shows the correlated water demand of one year. Users can choose different measures by clicking on the icons above the line chart and picking different years using the timeline slider on the top of the view. The other two maps on the right present the water supply and water scarcity respectively (Figure 3.12). The workflow of the GCAM visual analytics view is presented below (Figure 3.13).

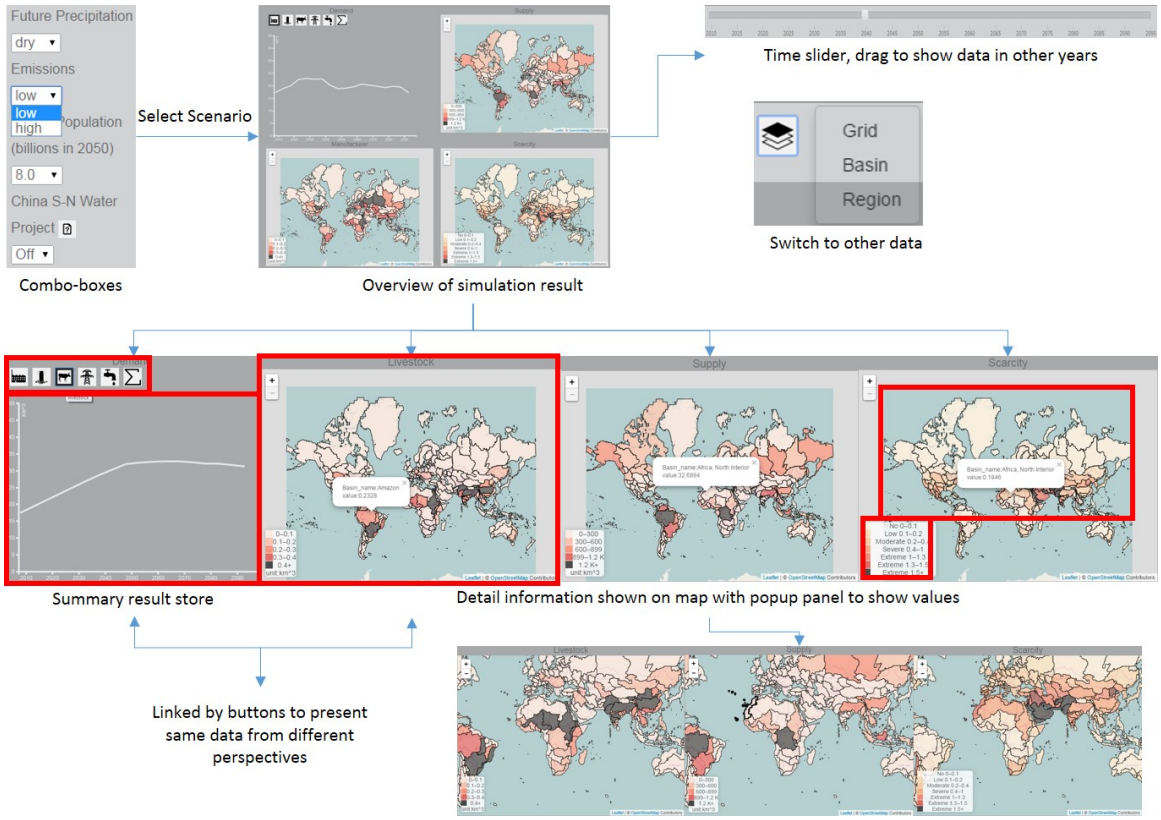

Figure 3.13: Workflow for the GCAM visual analytics view.

3.5.2 Similarity Analysis View

The similarity analysis view consists of four major sub-views: a geospatial view, a scenario similarity view, a timeline view, and a basin/region similarity view. The geospatial view allows users to highlight a specific area and apply the hierarchical clustering approach to those basins/regions in order to explore the impact of spatial variations and scales on scenario similarity with an interactive dendrogram. The basin similarity exploration view provides an interactive parallel coordinate plot for users to explore basin/region similarity results. The timeline view shows the average water scarcity level based on the selected scenarios over space and time.

I visualized the hierarchical clustering algorithm results through a dendrogram view (Figure 3.14). When a user selects basins/regions through a scalable rectan-

gle/polyline/polygon, the Euclidean distance among all sixty scenarios within the selected basins/regions will be calculated to produce a hierarchical tree. The scalable rectangle/polygon allows users to generate new hierarchical trees that can help users to understand the impact of the spatial scales and variations on scenario outputs. The leaf nodes represent each scenario. The parent nodes represent different categories (e.g., future precipitation, emissions) between two scenarios/two groups of scenarios. I use different colors to represent those different categories: red stands for future precipitation, yellow stands for emissions, green stands for population, and blue stands for China S-N Water Transportation Project. For example, the blue nodes in Figure 3.14 indicate that the two scenarios represented by two leaf nodes have the same future precipitation, emissions, and population, but one has the China S-N Water Transportation Project on and the other has the project off. Green nodes in Figure 3.14 indicate that two groups of scenarios represented by two child nodes have the same future precipitation, emissions, and the China S-N Water Transportation Project on and off, but different populations.

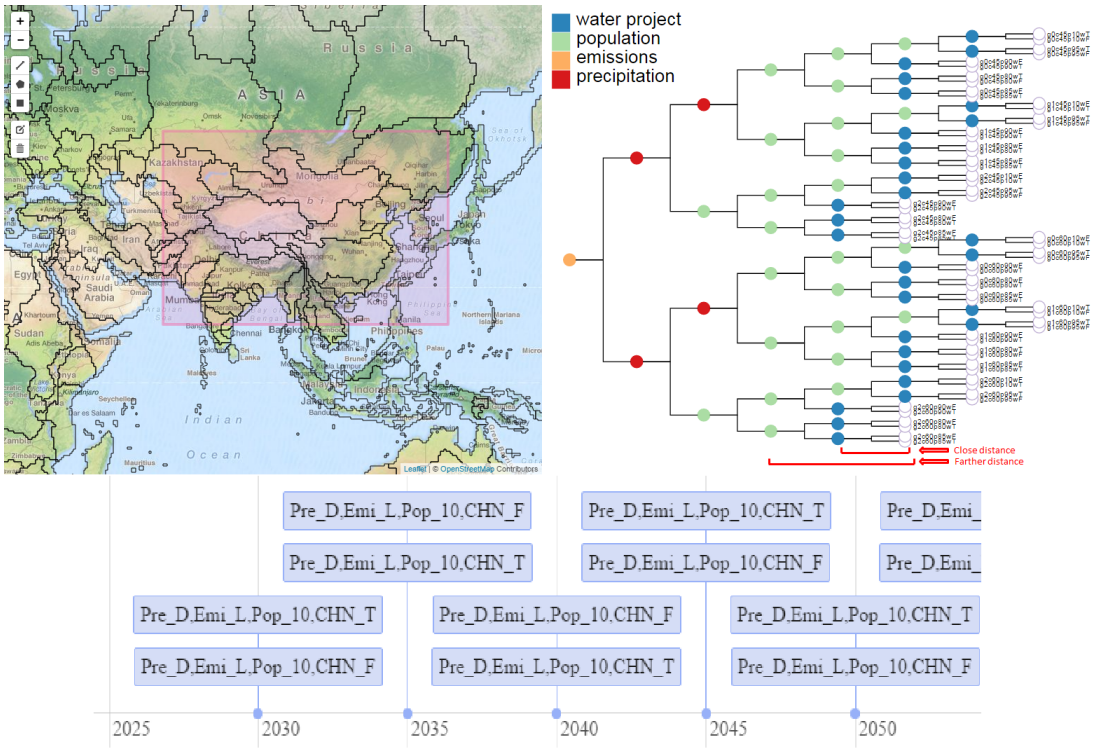

Figure 3.14: The map view on the top left, the dendrogram view on the top right, and the timeline view at the bottom. Map view allows users to select the basins and the dendrogram view shows the scenario similarity based on the selected basins. The timeline view shows that the average water scarcity values for the two scenarios within the selected basins in the map view are larger than a threshold value (0.4 in this example) since 2030. The two scenarios are selected through clicking one of the parent nodes (the enlarged blue node on the top) from the dendrogram view. One scenario is dry precipitation (Pre_D), low emission (Emi_L), a global population of 10 billion in 2050 (Pop_10), and the China South-North Water Transportation Project is on (CHN_T). The other scenario is dry precipitation, low emission, a global population of 10 billion in 2050, and the China South-North Water Transportation Project is off (CHN_F).

I propose a method to label the dendrogram with specific colors as described above. All scenarios include three future precipitations, two future emissions, five global populations, and two China South-North Water Transportation Projects. I use one vector to indicate one parameter with its possible values, so four parameters have four vectors in combination Figure 3.15 (a). Figure 3.15 (b) represents one scenario with a dry precipitation, a low emission, a global population of 10 billion in 2050, and the China South-North Water Transportation Project is on. I calculate

the difference between two scenarios (Figure 3.16) in order to label the parent nodes of the dendrogram view by the parameters with the highest absolute value:

$$Diff_P = \sum_{R \in P} AbsoluteValue(Cluster1_{P,R} - Cluster2_{P,R}) \tag{3.4}$$

Where R is the R^{th} record of parameter P.

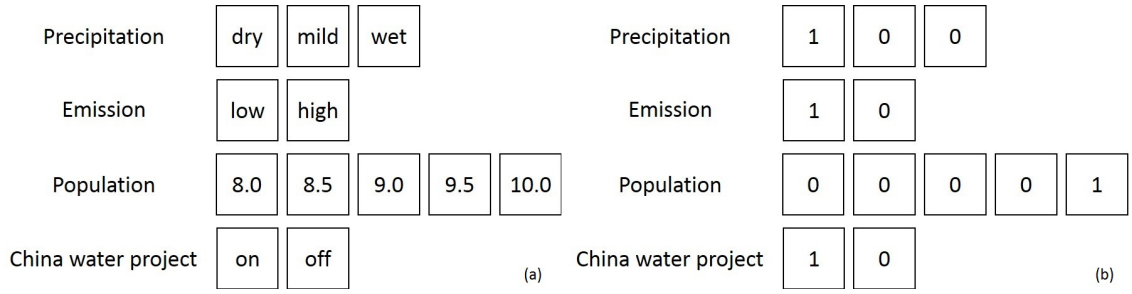

Figure 3.15: Vector representations of different scenarios. (a) All possible scenarios. (b) One scenario with a dry precipitation, a low emission, a global population of 10 billion in 2050, and the China South-North Water Transportation Project on.

For example, Figure 3.16 shows that the difference between the two scenarios is that one scenario has the China South-North Water Transportation Project on and the other has the project off, so I label the parent node from the dendrogram view in blue (Figure 3.17). When I label the parent nodes between two groups of scenarios, I need to merge two scenarios or two groups of scenarios into one new vector representation first Figure 3.15 (b), and then calculate the absolute difference between the two new vector representations (Figure 3.15 (a)).

The relative distance between parent nodes and the leaf nodes characterizes the dissimilarity level: farther distances indicate a high level of dissimilarity (Figure 3.14). Colored nodes at the lower level indicate which variables (e.g., China S-N Water Transportation Project) have a larger impact on the scenario similarity, whereas colored nodes at the higher level indicate which variables (e.g., future precipitation) have a smaller impact on the scenario similarity.

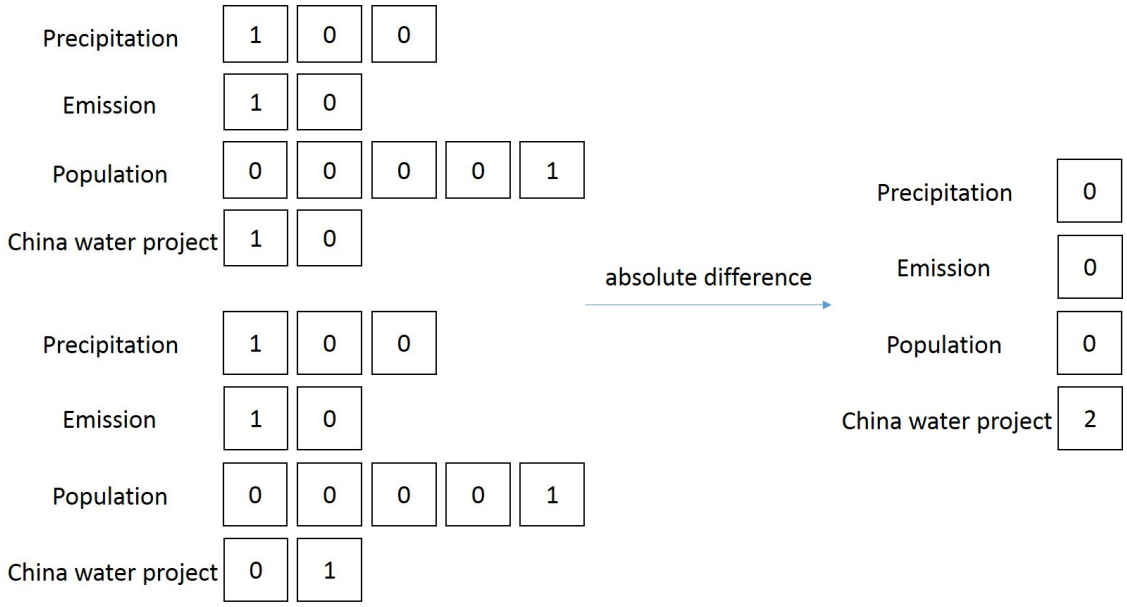

Figure 3.16: The difference calculation procedure for Dry Precipitation, Low Emission, 10.0 Billion Population and China S-N Water Project turned On (top) and Dry Precipitation, Low Emission, 10.0 Billion Population and China S-N Water Project turned Off (bottom).

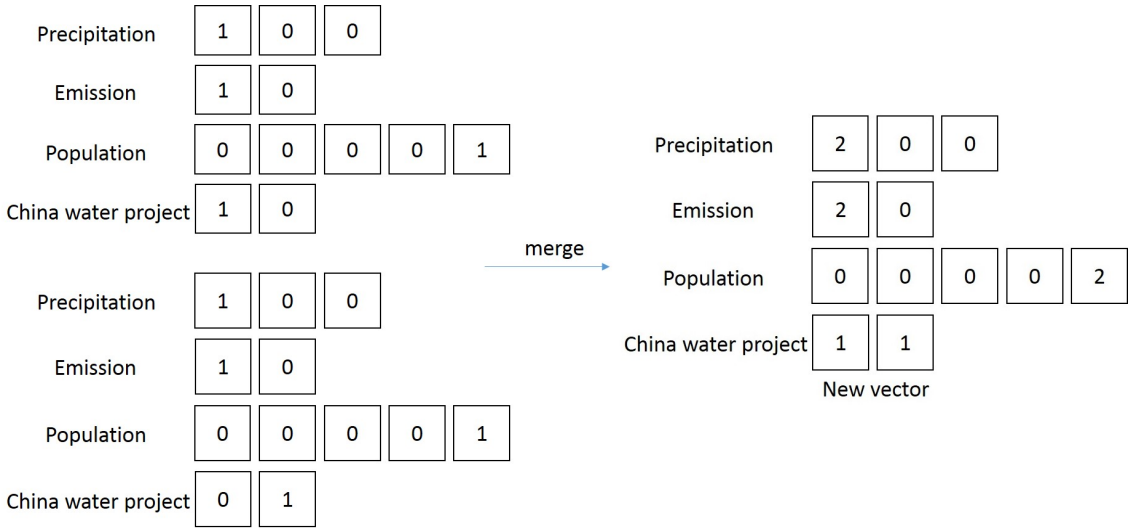

Figure 3.17: The merging procedure for Dry Precipitation, Low Emission, 10.0 Billion Population and China S-N Water Project turned On (top) and Dry Precipitation, Low Emission, 10.0 Billion Population and China S-N Water Project turned Off (bottom).

In addition to investigating the impact of spatial scales and variations on scenario similarity, this tool also allows the exploration of the average water scarcity level

within the selected basins/regions under different scenarios over time. Users can click any parent node from the dendrogram view (the enlarged blue node on the top in Figure 3.14). The scenarios under the selected parent nodes will be displayed on the timeline view when the average water scarcity value within the selected basins/regions is above a threshold value over time. The choice of the threshold value uses the water scarcity index (WSI) adopted from (Raskin *et al.* (1997)): no scarcity($0 \sim 0.1$), low scarcity ($0.1 \sim 0.2$), moderate scarcity ($0.2 \sim 0.4$), severe scarcity ($0.4 \sim 1$), and extreme scarcity (>1). The view helps users to understand the critical moments when the average water scarcity value within the selected areas and scenarios reaches a scarce level.

In terms of the basin similarity within one scenario, users can pick any scenario through the leaf node on the dendrogram view. According to the scenario, similarity computation will be applied to the selected basins/regions within a highlighted rectangle/polygon (Figure 3.18). Users can define the number of clusters and each cluster is assigned to one color. The results will be shown on both the map and the parallel coordinate plot (PCP). The two views allow the exploration of the basin similarity at both geographical and multidimensional attribute spaces. And for further analysis purposes, users can brush on PCP's axes to form a basin sub-group and rerun the clustering process, then analyze the new generated result and its distribution on the map view. The brushing function allows users to analyze a set of basins with higher similarity from the geospatial perspective (Figure 3.19).

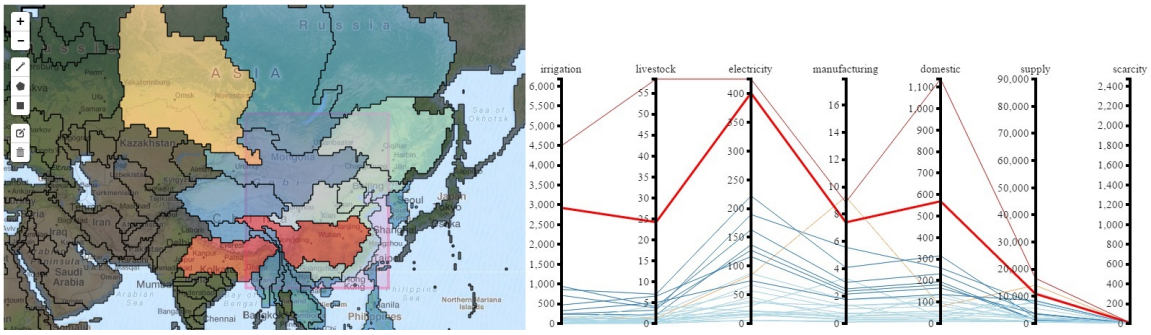

Figure 3.18: Basin similarity with a choropleth map and a parallel coordinate plot. Both display the basin similarity within a geographical boundary. The eight axes on the PCP represent six water demand variables (i.e., irrigation, livestock, electricity, manufacturing, and domestic), water supply, and water scarcity.

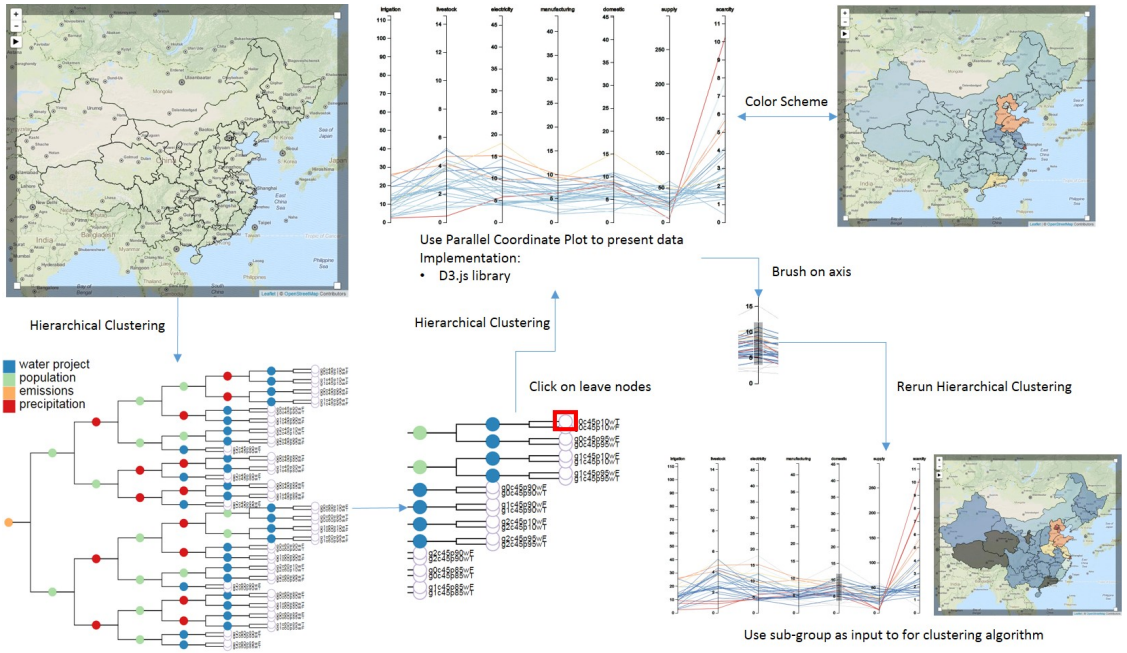

Figure 3.19: Workflow for the similarity analysis view

Chapter 4

CASE STUDY

I will now demonstrate how the GCAM visual analytics tool can help users explore and analyze the given global water data. I applied the tool to two case studies: China South−North Water Transportation Project and the impact of spatial variations and scales on scenario similarity. For China South−North Water Transportation Project, I analyze how this water project will impact on the future water scarcity of its donor basin (Yangtze River) and recipient basin (Ziya He). And for the impact of spatial variations and scales on scenario similarity, I explore the impact of spatial variations and scales on similarity analysis of climate scenarios varies at world, continental, and country scales.

4.1 China South−North Water Transportation Project

I apply the spatiotemporal view and scenario comparison view to explore the China South−North Water Transportation Project from the Yangtze River Basin to the Ziya He Basin. Figure 4.1 shows the spatial comparisons between Ziya He Basin in the China South−North Water Transportation Project. I pick the year 2030 with the scenario: dry future precipitation, low emissions, and 8 billion global population in the year of 2050. Compared to the left two figures, the right two figures show that the water supply and water scarcity appears to show some improvement when the China S−N project is implemented. The water supply increases from approximately 43 km^3 to approximately 72 km^3, whereas the water scarcity improves from extreme scarcity to severe scarcity. By exploring the detailed information from the spatial perspective, users can easily generate a hypothesis that the China South−North Water

Transportation Project improve the water scarcity of the Ziya He Basin. And temporal comparisons between models are needed to check whether this idea is correct or not over time. Figure 4.2 shows the temporal comparisons for water scarcity between the Yangtze River Basin and the Ziya He Basin in the China South−NorthWater Transporation Project from 2010 to 2095 under all possible combination scenarios among future precipitation, greenhouse gas emission, and population growth. All three sub figures exhibit consistent temporal patterns, no matter whether the China South−North Water Transportation Project carries on or not, the water scarcity becomes worse over time in the two basins. The water scarcity situations in Ziya He Basin improves significantly when the China South−North Water Transportation Project is implemented, whereas the China South−North Water Transportation Project does not dramatically worsen the water scarcity situations in Yangtze River Basin. These spatial and temporal views allow us to easily isolate the project's effects on scarcity over time in the simulation for both the donor and recipient basins.

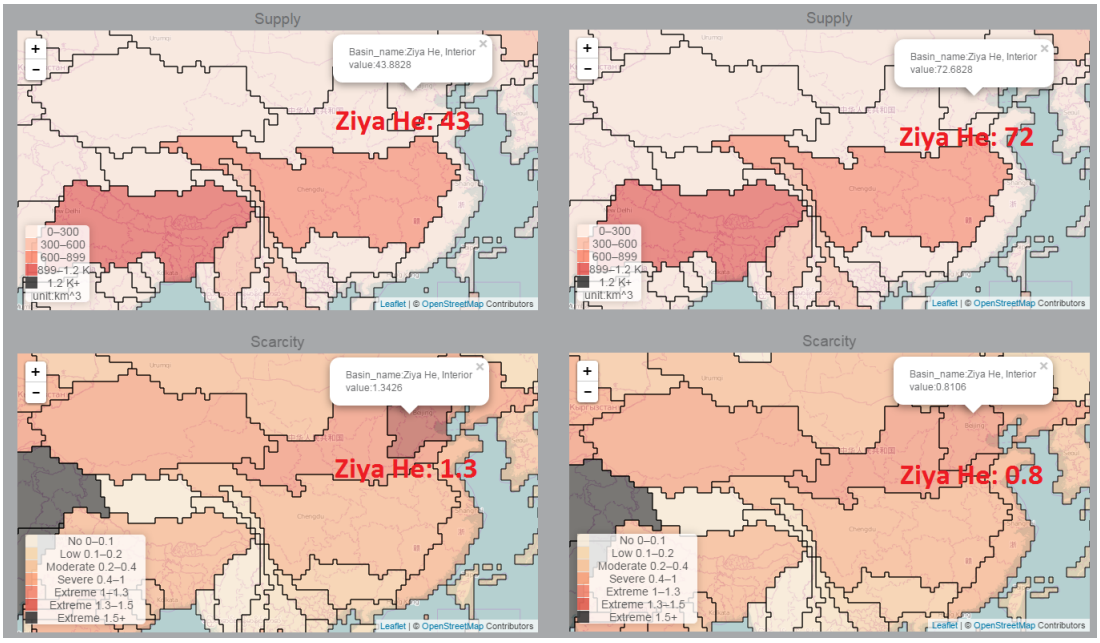

Figure 4.1: Spatial comparisons between water supply and water scarcity in 2030 in the China South-North Water Transportation Project. The left two figures shows results when the China South-North Water Transportation Project is off, and the right two figures show results when the China South-North Water Transportation Project is on.

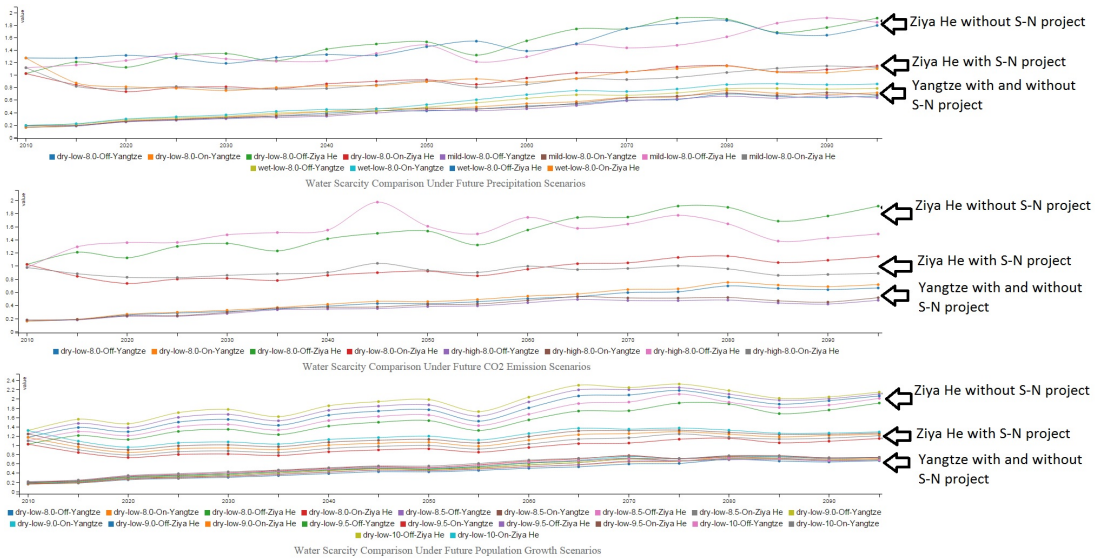

Figure 4.2: Temporal comparisons for water scarcity in the China South-North Water Transportation Project: water scarcity comparison under future precipitation scenarios, future greenhouse gas emission scenarios, and future population growth scenarios from top to bottom.

4.2 The Impact of Spatial Variations and Scales on Scenario Similarity

The GCAM model allows us to understand the complex interactions between climate models and human systems (e.g., population, climate policy). The output of the GCAM model consists of the water related data at both the basin and regional levels. Given that one goal of this paper is to explore the impact of spatial scales and variations on the GCAM model output, I compare the similarity analysis of climate scenarios at three different spatial scales: world, continental, and country scales. The world scale consists of 235 basins. The continent scale consists of five continents including Africa, Asia, America, Europe, and Australia at the basin level. The reason why I apply the basin data rather than regional data here is because the number of regions (62 including 31 divisions in China) is much smaller than the number of basins (235 in total) and most continents have no more than 5 regions. The reason is because the total number of basins within China is small (approximately 10) and some basins are beyond the geographical boundary of China.

4.2.1 World Scale

I first explore the scenario similarity based on all of the 235 basins across the world (Figure 4.3). From the dendrogram view on the right, I can see a clear hierarchical structure that the emission node in yellow is on the top of the tree, followed by the precipitation nodes in red, the population nodes in green, the water project nodes in blue, and all scenario leaf nodes. It implies that future emissions will have the largest impact on the scenario dissimilarity, followed by precipitation, population, and the China S-N Water Transportation Project. The results make sense for the following reasons. The China S-N Water Transportation Project only impacts two water basins that are the Yangtze River Basin and the Ziya He Basin, so it is supposed to have

the smallest impact on the scenario dissimilarity. The results tell us that, at the world scale, future climate policies on emissions make the biggest impact on climate scenario dissimilarities followed by water supply from GCMs, water demand from future population, and China S-N Water Transportation Project.

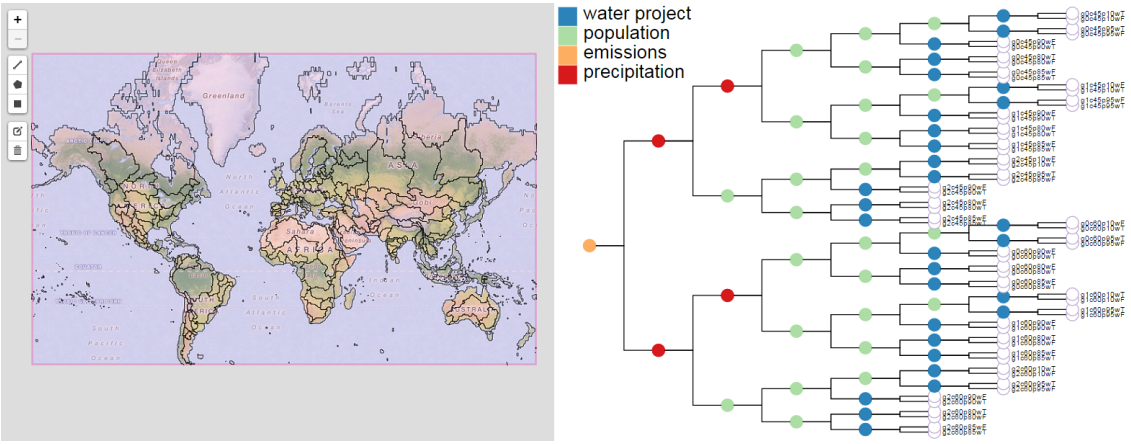

Figure 4.3: Scenario similarity based on all 235 water basins in the world.

4.2.2 Continental Scale and Variations

Figure 4.4 - 4.6 show the scenario similarity comparing continental variations in basins from Africa, Asia, and the Americas respectively. They all exhibit similar hierarchical structures as the world scale: the yellow emission node is on the top of the tree, followed by the red precipitation nodes, the green population nodes, the blue water project nodes, and all scenario leaf nodes. It also implies that the emission makes the biggest difference followed by the water supply from GCMs, water demand from future populations, and the China S-N Water Transportation Project

Figure 4.7 and Figure 4.8 show the scenario similarity based on basins from Europe and Australia respectively. From the dendrogram view on the right, I can see a different hierarchical structure from the rest of the world: emissions on the top, followed by population, future precipitation, and the China S-N Water Transportation

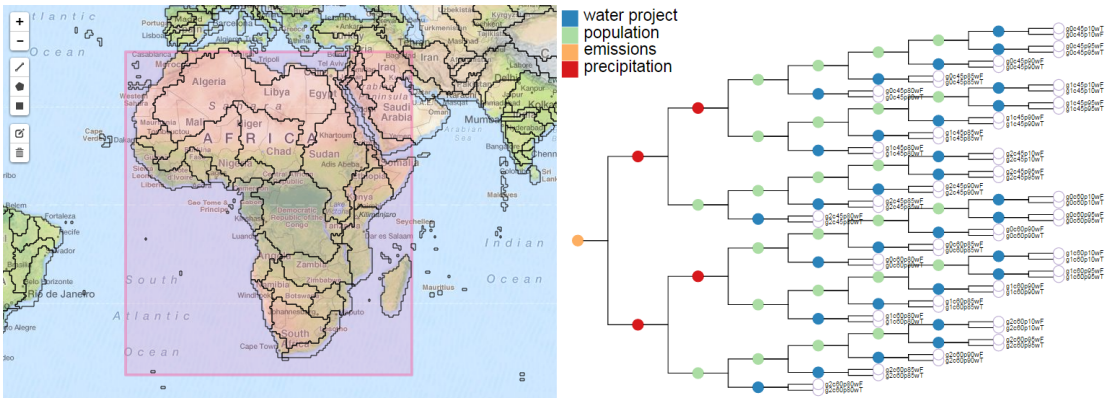

Figure 4.4: Scenario similarity based on water basins in Africa.

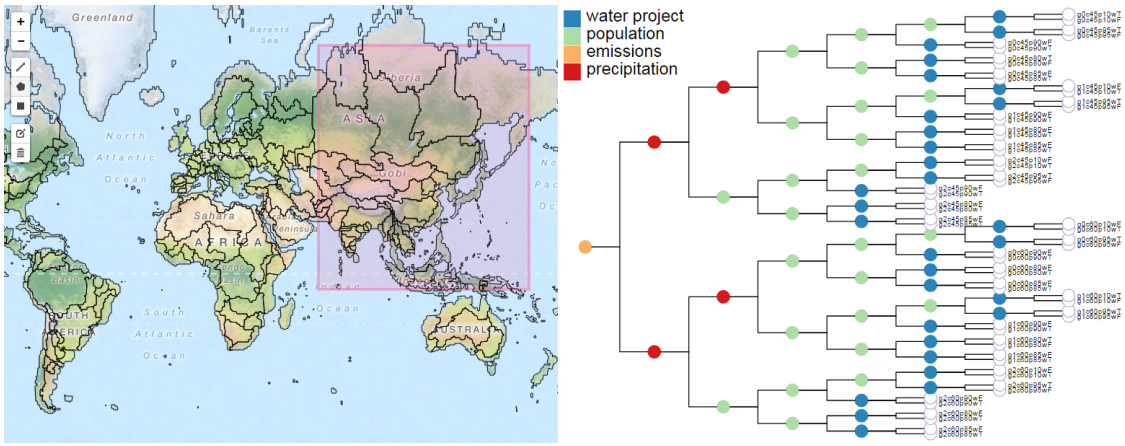

Figure 4.5: Scenario similarity based on water basins in Asia.

Project. The difference tells us that water demand from the population has a bigger impact on Europe and Australia than water supply from the GCMs compared to the rest of the world. Compared to the world scale in Figure 4.3, the differences tell us that spatial scales matter when I analyze scenario similarity, that is, the impact of water supply and demand on scenario similarity varies at different scales. Compared to the rest of the world at the continental scales in Figure 4.4, 4.5, and 4.6, the differences in Figure 4.7 and 4.8 tell us that the spatial variations matter when I analyze scenario similarity, that is, the impact of water supply and demand on scenario similarity varies over space.

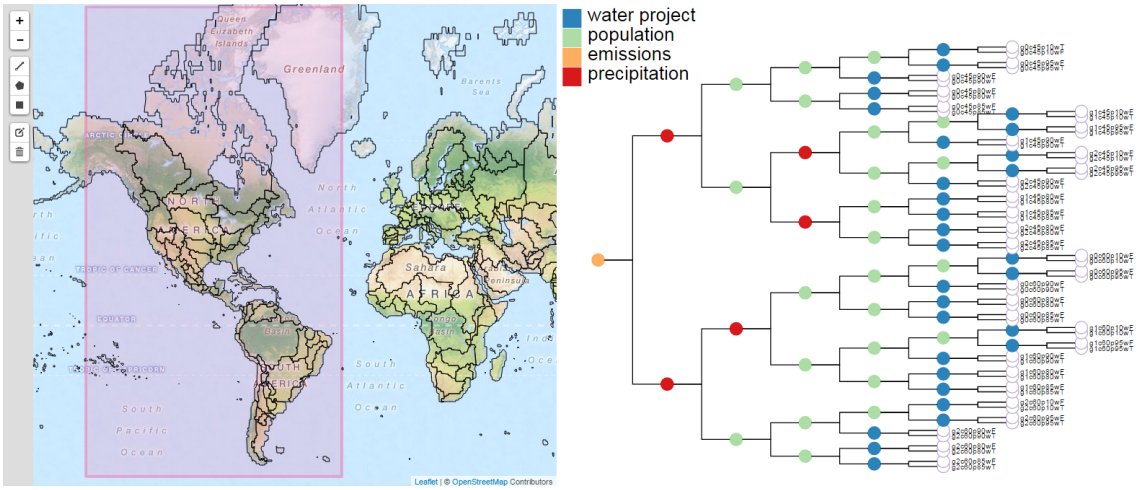

Figure 4.6: Scenario similarity based on water basins in America.

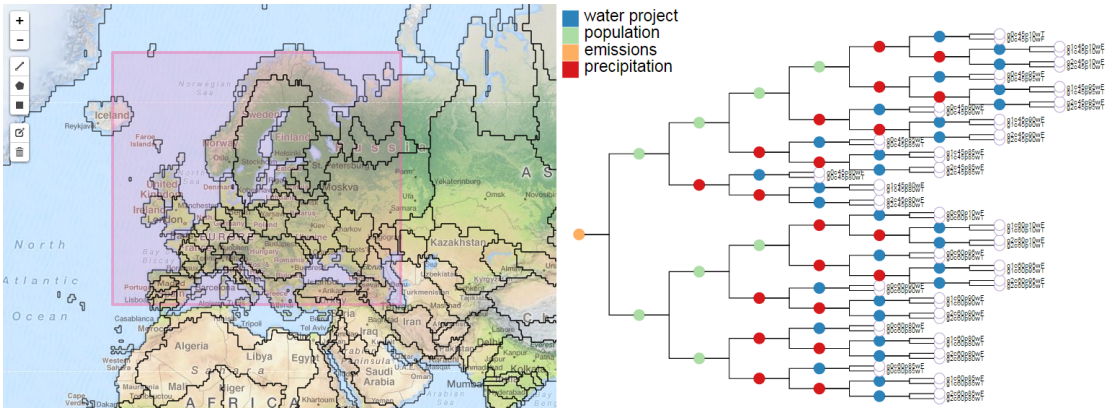

Figure 4.7: Scenario similarity based on water basins in Europe.

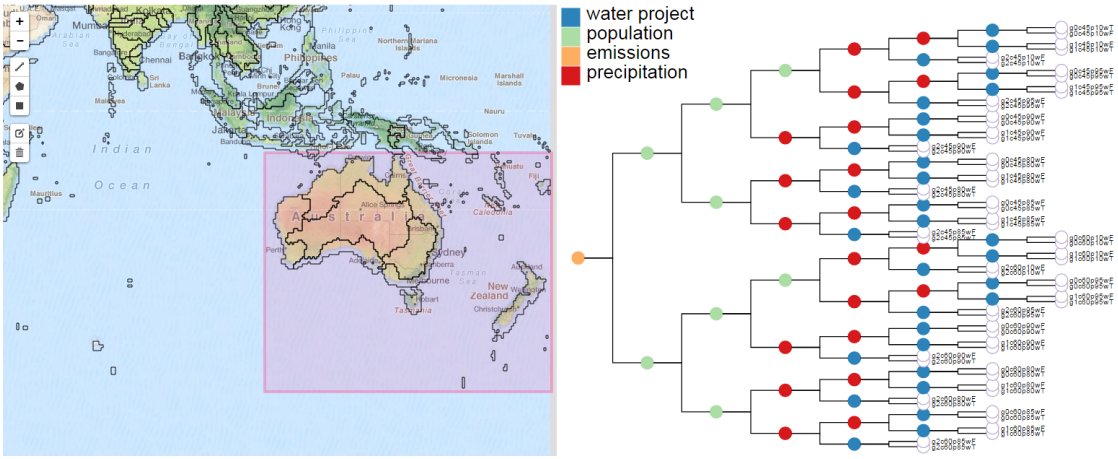

Figure 4.8: Scenario similarity based on water basins in Australia.

4.2.3 Country Scale

Figure 4.9 shows the scenario similarity based on the 31 divisions in China. The dendrogram view shows more complicated patterns in the hierarchical structure compared to the rest of the world (Figure 4.4, Figure 4.5, Figure 4.6, Figure 4.7, Figure 4.8): emissions are still on the top, followed the China S-N Water Transportation Project and mixed sequences of future precipitation and population. The timeline view in Figure 4.10 shows that the average scarcity values for the two scenarios in China are above a threshold value (0.4 in this case) since 2015. This indicates that China faces a severe water scarcity based on the two scenarios at a country level. To explore the potential reason of the water scarcity problem in the two scenarios, exploration of the divisions' data is needed to address which areas in China have a larger impact on the water scarcity overt time. I applied hierarchical clustering methods for comparing the similarity between all 31 division in China, and plot the result using the PCP and the map. The map and PCP views identify four different clusters based on the regional similarity: Shanghai in red, Beijing and Tianjing in yellow, Guangdong in dark blue, and the rest of China in light blue. From the PCP view, it shows that Shanghai has the highest scarcity level followed by Beijing and Tianjing, whereas Guangdong has the highest electricity and domestic demand. The rest of China has similar values across the seven water variables. Users can define a larger number of clusters to explore more details for the rest of China.

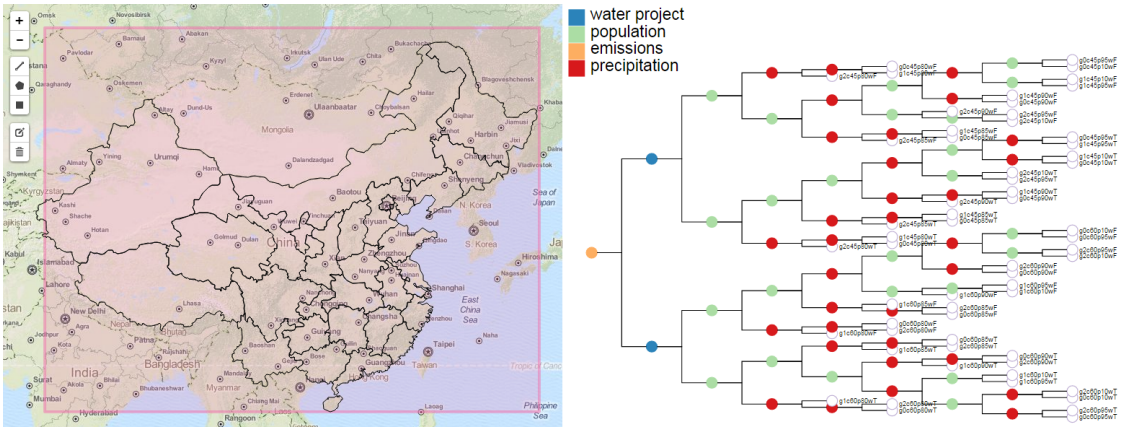

Figure 4.9: Scenario similarity based on the 31 divisions in China.

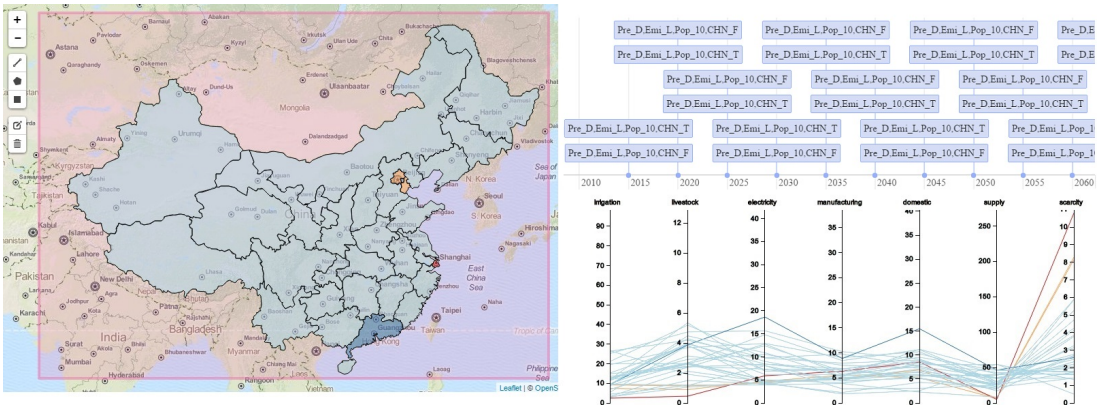

Figure 4.10: Regional similarity in China on both the map and PCP views with the defined cluster number as four. The timeline view shows that the average of scarcity values for the two scenarios (dry temperature, low climate change, and 10 billion global population with the China Project on and off) in China are above a threshold value (0.4 in this case) from 2015 to 2095.

Chapter 5

CONCLUSION AND FUTURE WORK

In this paper, I implement the first ever web-based geovisual analytics tool for GCAM to allow analysts to explore water supply, water demand, and water scarcity in the context of climate change and climate change policies. I demonstrate that such coupling allow analysts to explore and compare the impacts of different scenarios in both the spatial and temporal contexts. Furthermore, my tool also allows users to perform inter-comparison and similarity analysis of climate scenarios with GCAM models across the globe. The implemented GCAM visual analytics tool has integrated water availability models, which allow users to explore the interaction between climate models and human water systems. This tool implements a hierarchical clustering approach to perform similarity analysis among multiple scenarios and regional similarity within one scenario. This tool also allows users to explore the average of scarcity values according to the selected scenarios over space and time. Based on the GCAM models, this tool demonstrates the impact of spatial variations and scales on similarity analysis of climate scenarios varies at world, continental, and country scales.

There are many extensions and work that are worth further explorations. I will continue this work in terms of eight possible extensions of this tool. Given a close collaboration with model developers and analysts from the GCAM team, I will continuously apply the feedback from them as the user-centered design guidelines for future implementation. More clustering approaches to perform inter-comparison of scenario analysis will be implemented in order to develop comprehensive metrics to quantify similarity and dissimilarity among different scenarios. Given that GCAM simula-

tions take place externally, the next step is to couple the visualization tool with the simulation models into an integrated system. Given that this tool uses the NetCDF structure, which is the universal data model in climate research area, this tool can be easily extended for similar scenario comparison beyond water system. Given that this approach provides similarity analysis for scenario results based on climate computational models with varied parameters, I will add more features and apply my tool to solve similar problems related to different computational models which involve similar spatiotemporal data structures. Given the large number of output variables for a large number of sectors, I will expand my tool with more methods for feature exploration, comparison, and configuration (with the help of these functions, users can gain an overview first, eliminate outliers, select groups of variables, then apply the configuration to the similarity comparison process). To improve insight into the complex blackbox model and help users understand unexpected difference in models, I will design and develop more robust features including visual components such as self-organizing map (SOM) with various interactive functions which allow users to communicate with GCAM from all perspectives and get feedback with detail information. I'll iteratively improve the user interface by adding themes ,color schemes, also animation to make the front-end more users friendly.

REFERENCES

Andrienko, N. and G. Andrienko, "A visual analytics framework for spatio-temporal analysis and modelling", Data Mining and Knowledge Discovery **27**, 1, 55–83 (2013).

Beer, C., M. Reichstein, E. Tomelleri, P. Ciais, M. Jung, N. Carvalhais, C. Rödenbeck, M. A. Arain, D. Baldocchi, G. B. Bonan *et al.*, "Terrestrial gross carbon dioxide uptake: global distribution and covariation with climate", Science **329**, 5993, 834–838 (2010).

Bostock, M., "D3. js", Data Driven Documents (2012).

Brus, J., V. Voženílek and S. Popelka, "An assessment of quantitative uncertainty visualization methods for interpolated meteorological data", in "Computational Science and Its Applications–ICCSA 2013", pp. 166–178 (Springer, 2013).

Buckley, S. and C. An, "Supply chain simulation", in "Supply Chain Management on Demand", pp. 17–35 (Springer, 2005).

Buja, A., D. Cook and D. F. Swayne, "Interactive high-dimensional data visualization", Journal of Computational and Graphical Statistics **5**, 1, 78–99 (1996).

Burbeck, S., "Applications programming in smalltalk-80 (tm): How to use Model-View-Controller (MVC)", Smalltalk-80 v2 **5** (1992).

Burch, S., S. R. Sheppard, A. Shaw and D. Flanders, "Planning for climate change in a flood-prone community: municipal barriers to policy action and the use of visualizations as decision-support tools", Journal of Flood Risk Management **3**, 2, 126–139 (2010).

Clarke, L., J. Edmonds, H. Jacoby, H. Pitcher, J. Reilly and R. Richels, "Scenarios of greenhouse gas emissions and atmospheric concentrations", US Department of Energy Publications p. 6 (2007a).

Clarke, L., J. Lurz, M. Wise, J. Edmonds, S. Kim, S. Smith and H. Pitcher, "Model documentation for the minicam climate change science program stabilization scenarios: Ccsp product 2.1 a", Pacific Northwest National Laboratory, PNNL-16735 (2007b).

Dasgupta, A., J. Poco, Y. Wei, R. Cook, E. Bertini and C. Silva, "Bridging theory with practice: An exploratory study of visualization use and design for climate model comparison", Visualization and Computer Graphics, IEEE (2015).

Deza, M. M. and E. Deza, *Encyclopedia of distances* (Springer, Berlin, 2009).

Dockerty, T., A. Lovett, G. Sünnenberg, K. Appleton and M. Parry, "Visualising the potential impacts of climate change on rural landscapes", Environment and Urban Systems, Computers **29**, 3, 297–320 (2005).

ESRI, "Esri shapefile technical description", (1998).

Everitt, B. S. and A. Skrondal, "The cambridge dictionary of statistics", Cambridge: Cambridge (2002).

Foley, J. D., A. Van Dam, S. K. Feiner, J. F. Hughes and R. L. Phillips, *Introduction to computer graphics*, vol. 55 (Addison-Wesley Reading, 1994).

Furrer, R., R. Knutti, S. Sain, D. Nychka and G. Meehl, "Spatial patterns of probabilistic temperature change projections from a multivariate Bayesian analysis", Geophysical Research Letters **34**, 6 (2007).

Hejazi, M., J. Edmonds, L. Clarke, P. Kyle, E. Davies, V. Chaturvedi, J. Eom, M. Wise, P. Patel and K. Calvin, "Integrated assessment of global water scarcity over the 21st century–part 2: Climate change mitigation policies", Hydrol. Earth Syst. Sci. Discuss **10**, 3, 3383–3425 (2013a).

Hejazi, M., J. Edmonds, L. Clarke, P. Kyle, E. Davies, V. Chaturvedi, M. Wise, P. Patel, J. Eom and K. Calvin, "Integrated assessment of global water scarcity over the 21st century–part 1: Global water supply and demand under extreme radiative forcing", Hydrol. Earth Syst. Sci. Discuss **10**, 3, 3327–3381 (2013b).

Hejazi, M., J. Edmonds, L. Clarke, P. Kyle, E. Davies, V. Chaturvedi, M. Wise, P. Patel, J. Eom, K. Calvin *et al.*, "Long-term global water projections using six socioeconomic scenarios in an integrated assessment modeling framework", Technological Forecasting and Social Change **81**, 205–226 (2014a).

Hejazi, M. I., J. Edmonds, L. Clarke, P. Kyle, E. Davies, V. Chaturvedi, M. Wise, P. Patel, J. Eom and K. Calvin, "Integrated assessment of global water scarcity over the 21st century under multiple climate change mitigation policies", Hydrology and Earth System Sciences **18**, 8, 2859–2883 (2014b).

Helbig, C., H.-S. Bauer, K. Rink, V. Wulfmeyer, M. Frank and O. Kolditz, "Concept and workflow for 3d visualization of atmospheric data in a virtual reality environment for analytical approaches", Environmental Earth Sciences **72**, 10, 3767–3780 (2014).

Houghton, A., N. Prudent, J. E. Scott, R. Wade and G. Luber, "Climate change-related vulnerabilities and local environmental public health tracking through gemss: A web-based visualization tool", Applied Geography **33**, 36–44 (2012).

Hubmann-Haidvogel, A., A. M. Brasoveanu, A. Scharl, M. Sabou and S. Gindl, "Visualizing contextual and dynamic features of micropost streams", 2nd Workshop on Making Sense of Microposts (MSM-2012), 21st International World Wide Web Conference (2012).

Hubmann-Haidvogel, A., A. Scharl and A. Weichselbraun, "Multiple coordinated views for searching and navigating web content repositories", Information Sciences **179**, 12, 1813–1821 (2009).

Huntzinger, D., C. Schwalm, A. Michalak, K. Schaefer, A. King, Y. Wei, A. Jacobson, S. Liu, R. Cook, W. Post *et al.*, "The north american carbon program multi-scale synthesis and terrestrial model intercomparison project–part 1: overview and experimental design", Geoscientific Model Development **6**, 6, 2121–2133 (2013).

Ismail, M., I. J. Musa, A. Salisu, S. Zubairu, I. Kim, I. Musa, I. Ibrahim and Z. Muhammed, "Visualising the relationship between energy use and climate change", Current Advances in Environmental Science (CAES) **2**, 1, 1–10 (2014).

Jin, H. and D. Guo, "Understanding climate change patterns with multivariate geovisualization", in "Data Mining Workshops, IEEE International Conference on, 2009. ICDMW'09.", pp. 217–222 (IEEE, 2009).

Kaye, N., A. Hartley and D. Hemming, "Mapping the climate: guidance on appropriate techniques to map climate variables and their uncertainty", Geoscientific Model Development **5**, 1, 245–256 (2012).

Kehrer, J., F. Ladstadter, P. Muigg, H. Doleisch, A. Steiner and H. Hauser, "Hypothesis generation in climate research with interactive visual data exploration", Visualization and Computer Graphics, IEEE Transactions on **14**, 6, 1579–1586 (2008).

Kim, S. H., J. Edmonds, J. Lurz, S. J. Smith and M. Wise, "The objects framework for integrated assessment: Hybrid modeling of transportation", The Energy Journal pp. 63–91 (2006).

Knutti, R., "Should we believe model predictions of future climate change?", Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences **366**, 1885, 4647–4664 (2008).

Knutti, R., D. Masson and A. Gettelman, "Climate model genealogy: Generation cmip5 and how we got there", Geophysical Research Letters **40**, 6, 1194–1199 (2013).

Kohlhammer, J., K. Nazemi, T. Ruppert and D. Burkhardt, "Toward visualization in policy modeling", Computer Graphics and Applications, IEEE **32**, 5, 84–89 (2012).

Kothur, P., M. Sips, H. Dobslaw and D. Dransch, "Visual analytics for comparison of ocean model output with reference data: Detecting and analyzing geophysical processes using clustering ensembles", Visualization and Computer Graphics, IEEE Transactions on **20**, 12, 1893–1902 (2014).

Ladstädter, F., A. K. Steiner, B. C. Lackner, B. Pirscher, G. Kirchengast, J. Kehrer, H. Hauser, P. Muigg and H. Doleisch, "Exploration of climate data using interactive visualization*", Journal of Atmospheric and Oceanic Technology **27**, 4, 667–679 (2010).

Luo, W., *Geovisual Analytics Approaches for the Integration of Geography and Social Network Contexts*, Ph.D. thesis, The Pennsylvania State University (2014).

Maciejewski, R., P. Livengood, S. Rudolph, T. F. Collins, D. S. Ebert, R. T. Brigantic, C. D. Corley, G. A. Muller and S. W. Sanders, "A pandemic influenza modeling and visualization tool", Journal of Visual Languages & Computing **22**, 4, 268–278 (2011).

Malik, A., R. Maciejewski, S. Towers, S. McCullough and D. S. Ebert, "Proactive spatiotemporal resource allocation and predictive visual analytics for community policing and law enforcement", Visualization and Computer Graphics, IEEE Transactions on **20**, 12, 1863–1872 (2014).

Masson, D. and R. Knutti, "Climate model genealogy", Geophysical Research Letters **38**, 8 (2011).

Matković, K., D. Gračanin, M. Jelović, A. Ammer, A. Lež and H. Hauser, "Interactive visual analysis of multiple simulation runs using the simulation model view: Understanding and tuning of an electronic unit injector", Visualization and Computer Graphics, IEEE Transactions on **16**, 6, 1449–1457 (2010).

Nocke, T., M. Flechsig and U. Böhm, "Visual exploration and evaluation of climate-related simulation data", in "Simulation Conference, 2007 Winter", pp. 703–711 (IEEE, 2007).

Nocke, T., T. Sterzel, M. Böttinger and M. Wrobel, "Visualization of climate and climate change data: An overview", Digital earth summit on geoinformatics pp. 226–232 (2008).

Pang-Ning, T., M. Steinbach, V. Kumar *et al.*, "Introduction to data mining", in "Library of Congress", p. 74 (2006).

Parker, W. S., "Ensemble modeling, uncertainty and robust predictions", Wiley Interdisciplinary Reviews: Climate Change **4**, 3, 213–223 (2013).

Pennell, C. and T. Reichler, "On the effective number of climate models", Journal of Climate **24**, 9, 2358–2367 (2011).

Poco, J., A. Dasgupta, Y. Wei, W. Hargrove, C. Schwalm, R. Cook, E. Bertini and C. Silva, "Similarityexplorer: A visual inter-comparison tool for multifaceted climate data", in "Computer Graphics Forum", vol. 33, pp. 341–350 (Wiley Online Library, 2014a).

Poco, J., A. Dasgupta, Y. Wei, W. Hargrove, C. R. Schwalm, D. N. Huntzinger, R. Cook, E. Bertini and C. T. Silva, "Visual reconciliation of alternative similarity spaces in climate modeling", Visualization and Computer Graphics, IEEE Transactions on **20**, 12, 1923–1932 (2014b).

Potter, K., A. Wilson, P.-T. Bremer, D. Williams, C. Doutriaux, V. Pascucci and C. Johhson, "Visualization of uncertainty and ensemble data: Exploration of climate modeling and weather forecast data with integrated visus-cdat systems", in "Journal of Physics: Conference Series", vol. 180, p. 012089 (IOP Publishing, 2009a).

Potter, K., A. Wilson, P.-T. Bremer, D. Williams, C. Doutriaux, V. Pascucci and C. R. Johnson, "Ensemble-vis: A framework for the statistical visualization of ensemble data", in "IEEE International Conference on Data Mining Workshops", pp. 233–240 (IEEE, 2009b).

Raskin, P., P. Gleick, P. Kirshen, G. Pontius and K. Strzepek, *Water futures: assessment of long-range patterns and problems. Comprehensive assessment of the freshwater resources of the world* (SEI, 1997).

Reenskaug, T. and J. O. Coplien, "The dci architecture: A new vision of object-oriented programming", An article starting a new blog:(14pp) http://www. artima. com/articles/dci_vision. html (2009).

Ribicic, H., J. Waser, R. Fuchs, G. Bloschl and E. Groller, "Visual analysis and steering of flooding simulations", Visualization and Computer Graphics, IEEE Transactions on **19**, 6, 1062–1075 (2013).

Sanyal, J., S. Zhang, J. Dyer, A. Mercer, P. Amburn and R. J. Moorhead, "Noodles: A tool for visualization of numerical weather model ensemble uncertainty", Visualization and Computer Graphics, IEEE Transactions on **16**, 6, 1421–1430 (2010).

Shneiderman, B., "The eyes have it: A task by data type taxonomy for information visualizations", in "Visual Languages, IEEE", pp. 336–343 (IEEE, 1996).

Steed, C. A., D. M. Ricciuto, G. Shipman, B. Smith, P. E. Thornton, D. Wang, X. Shi and D. N. Williams, "Big data visual analytics for exploratory earth system simulation analysis", Computers & Geosciences **61**, 71–82 (2013).

Steed, C. A., G. Shipman, P. Thornton, D. Ricciuto, D. Erickson and M. Branstetter, "Practical application of parallel coordinates for climate model analysis", Procedia Computer Science **9**, 877–886 (2012).

Sun, X., S. Shen, G. G. Leptoukh, P. Wang, L. Di and M. Lu, "Development of a web-based visualization platform for climate research using Google Earth", Computers & Geosciences **47**, 160–168 (2012).

Szekely, G. J. and M. L. Rizzo, "Hierarchical clustering via joint between-within distances: Extending ward's minimum variance method", Journal of Classification **22**, 2, 151–183 (2005).

Tebaldi, C. and R. Knutti, "The use of the multi-model ensemble in probabilistic climate projections", Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences **365**, 1857, 2053–2075 (2007).

Tebaldi, C., R. L. Smith, D. Nychka and L. O. Mearns, "Quantifying uncertainty in projections of regional climate change: A Bayesian approach to the analysis of multimodel ensembles", Journal of Climate **18**, 10, 1524–1540 (2005).

Tomko, M., P. Greenwood, M. Sarwar, L. Morandini, R. Stimson, C. Bayliss, G. Galang, M. Nino-Ruiz, W. Voorsluys, I. Widjaja *et al.*, "The design of a flexible web-based analytical platform for urban research", in "Proceedings of the 20th International Conference on Advances in Geographic Information Systems", pp. 369–375 (ACM, 2012).

Vörösmarty, C. J., P. Green, J. Salisbury and R. B. Lammers, "Global water resources: vulnerability from climate change and population growth", Science **289**, 5477, 284–288 (2000).

Vörösmarty, C. J., P. McIntyre, M. O. Gessner, D. Dudgeon, A. Prusevich, P. Green, S. Glidden, S. E. Bunn, C. A. Sullivan, C. R. Liermann *et al.*, "Global threats to human water security and river biodiversity", Nature **467**, 7315, 555–561 (2010).

Wang, C., H. Yu and K.-L. Ma, "Importance-driven time-varying data visualization", Visualization and Computer Graphics, IEEE Transactions on **14**, 6, 1547–1554 (2008).

Waser, J., R. Fuchs, H. Ribičić, B. Schindler, G. Blöschl and M. E. Gröller, "World lines", Visualization and Computer Graphics, IEEE Transactions on **16**, 6, 1458–1467 (2010).

Williams, D., D. Charles, P. John, W. Sean, S. Galen, M. Ross and S. Chad, "The ultra-scale visualization climate data analysis tools (uv-cdat): Data analysis and visualization for geoscience data", IEEE Computer (2014).

Yokohata, T., J. D. Annan, M. Collins, C. S. Jackson, M. Tobis, M. J. Webb and J. C. Hargreaves, "Reliability of multi-model and structurally different single-model ensembles", Climate Dynamics **39**, 3-4, 599–616 (2012).